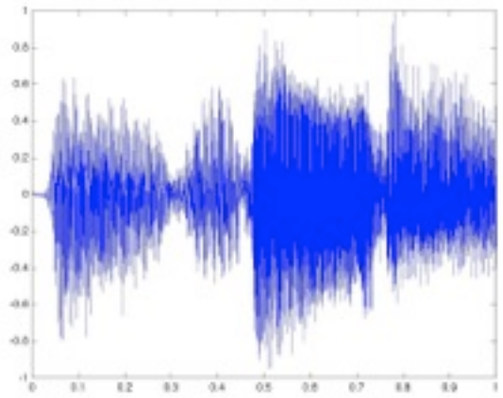
Deep Network Geometry of Timbre

Vincent Lostanlen, Stéphane Mallat
Joan Bruna, Joaquim Andèn

École Normale Supérieure
www.di.ens.fr/data

High Dimensional Classification

$d = 10^4/s$



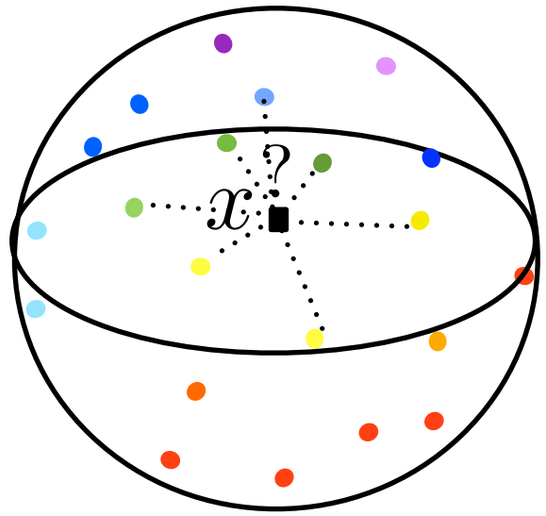
- Audio signal $x = (x(1), ..., x(d)) \in \mathbb{R}^d$:

- **Classification:** estimate a class label $f(x)$ given n sample values $\{x_i\,,\, y_i = f(x_i)\}_{i \leq n}$

- Why is it so difficult ?

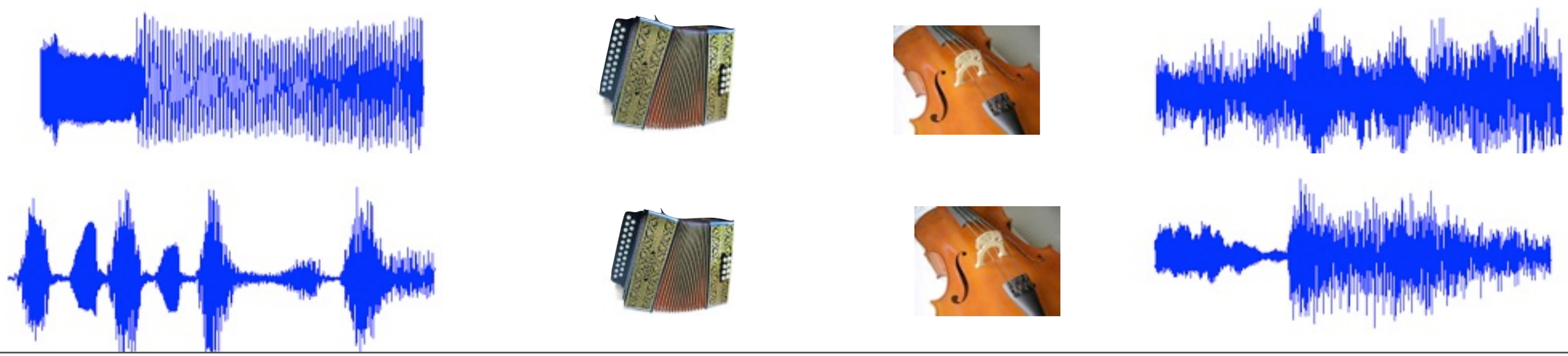
Curse of Dimensionality

- $f(x)$ can be approximated from examples $\{x_i\,,\, f(x_i)\}_i$ by local interpolation if f is regular and there are close examples:



- Need ϵ^{-d} points to cover $[0,1]^d$ at a Euclidean distance ϵ

$$\Rightarrow \|x - x_i\| \text{ is always large}$$



Euclidean Embedding

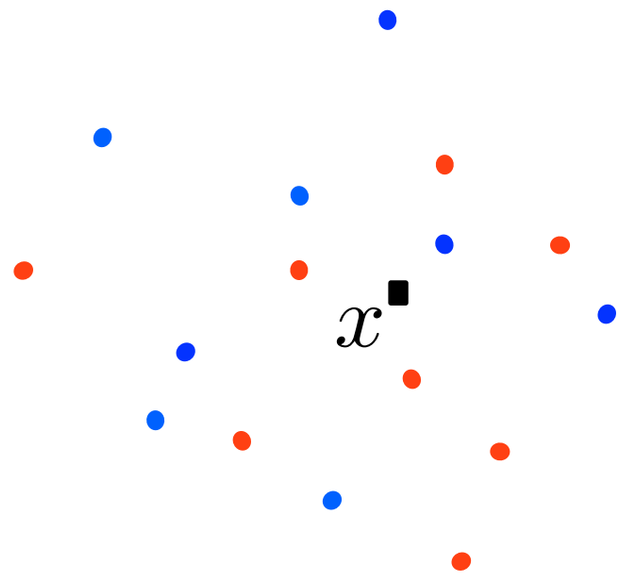
Data: $x \in \mathbb{R}^d$

$\|x - x'\|$: non-informative

Representation

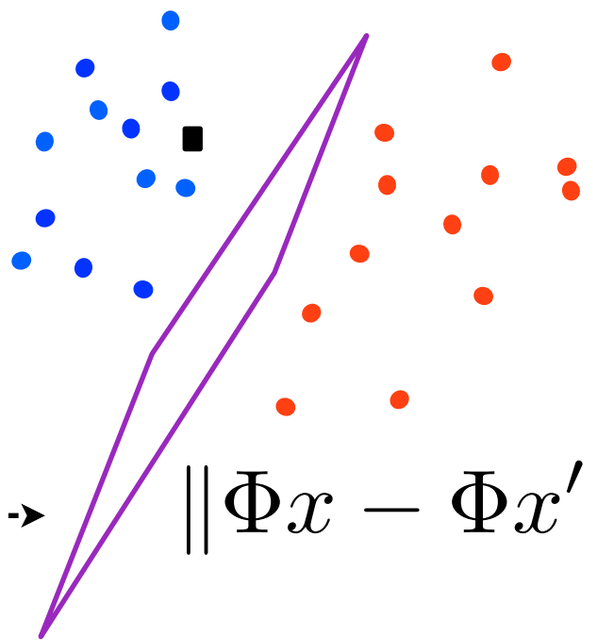
$\Phi x \in \mathcal{H}$

Linear Classifier



Intelligence

Φ



"Similarity" metric: $\Delta(x, x')$ $\longleftrightarrow$ $\|\Phi x - \Phi x'\|$

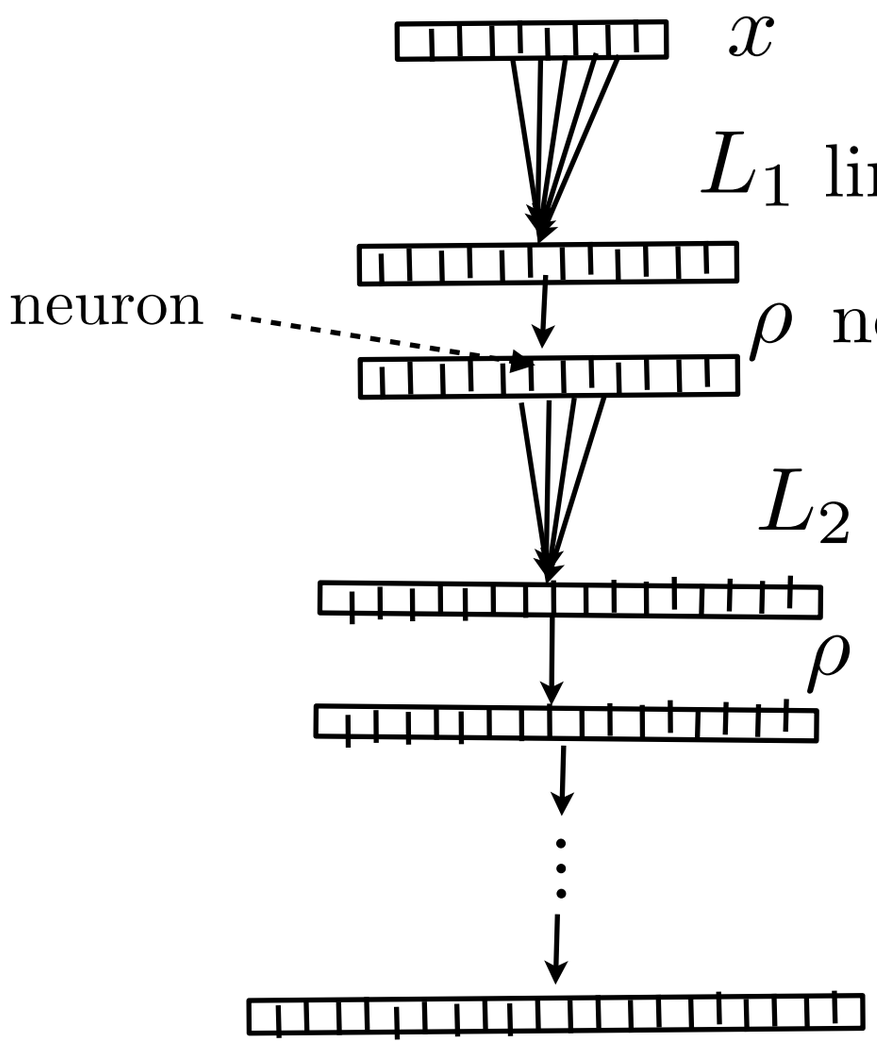
Equivalent Euclidean metric:

$$C_1 \|\Phi x - \Phi x'\| \leq \Delta(x, x') \leq C_2 \|\Phi x - \Phi x'\|$$

How to define Φ ?

Deep Convolution Neworks

- The revival of an old (1950) idea: *Y. LeCun*



: Wavelets !

Optimize the L_k with architecture conditions: over 10^9 paramet

Exceptional results for *images, speech, bio-data* classification.

Products by FaceBook, IBM, Google, Microsoft, Yahoo...

Why does it work so well ?

Overview of Questions

- How to build audio signal representations for classification ?
- Why are deep neural networks so efficient ?
- Why do wavelets appear in the cochlea and in most classifiers ?
- Why non-linearities ?

A Geometric Approach to Timbre

Geometric Representation

- What geometry ? $\longrightarrow_t$ *quite poor...*

- **Invariance** to translations $x_c(t) = x(t-c)$

$$\forall c \in \mathbf{R} \quad , \quad \Phi(x_c) = \Phi(x) \ .$$

- **Stability** to deformations $x_\tau(t) = x(t-\tau(t))$

small deformations of $x \implies$ small modifications of $\Phi(x)$

$$\forall \tau \quad , \quad \|\Phi(x_\tau) - \Phi(x)\| \leq C \underbrace{\sup_t |\tau'(t)|}_{\text{deformation size}} \|x\| \ .$$

- **Preserve information**

Fourier Translation Invariance

- Fourier transform $\hat{x}(\omega) = \int x(t)\, e^{-i\omega t}\, dt$ invariance:

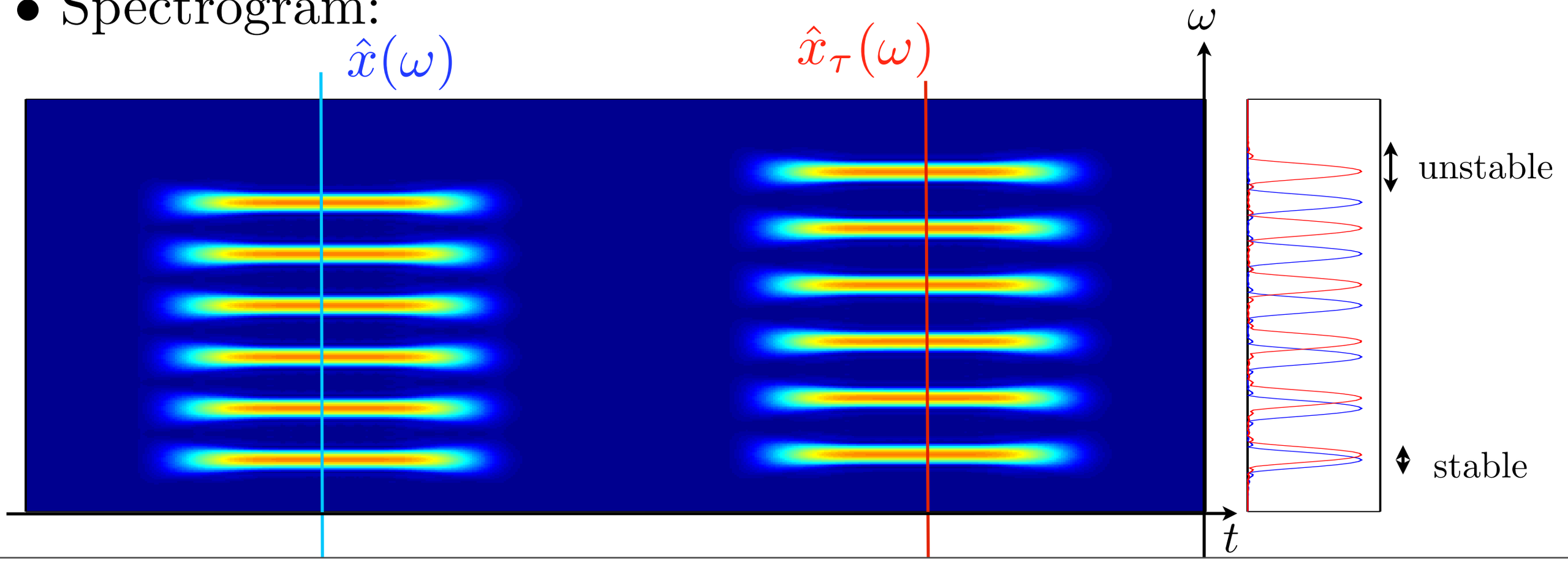
$$\text{if } x_c(t) = x(t-c) \text{ then } |\widehat{x_c}(\omega)| = |\hat{x}(\omega)|$$

- Instabilites to small deformations $x_\tau(t) = x(t - \tau(t))$:

$$\big|\, |\hat{x}_\tau(\omega)| - |\,\hat{x}(\omega)|\,\big| \quad \text{is big at high frequencies}$$

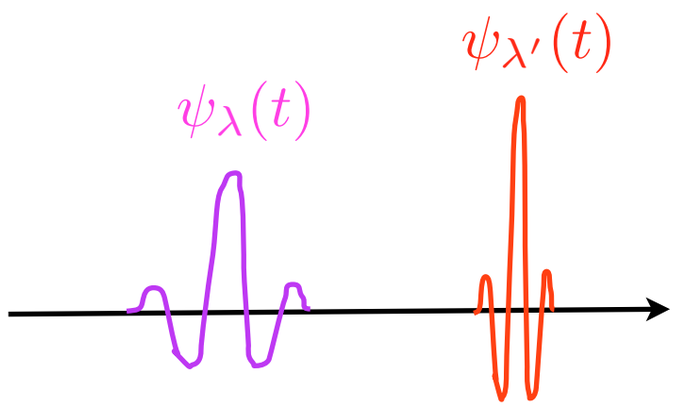
$$\tau(t) = \epsilon\, t$$

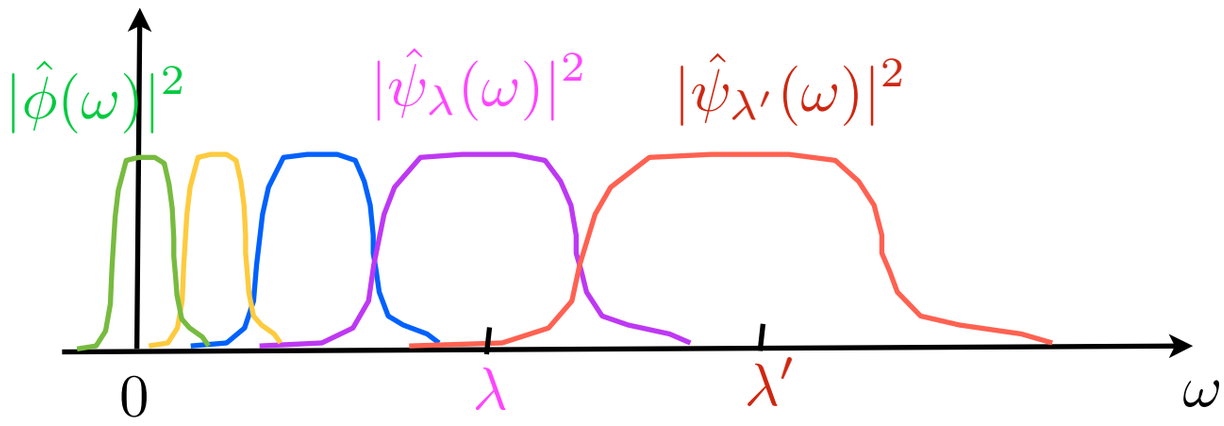
- Spectrogram:



Wavelet Transform

- Dilated wavelets: $\psi_\lambda(t) = 2^{-j/Q}\,\psi(2^{-j/Q}t)$ with $\lambda = 2^{-j/Q}$





- Choice of Q: sparsity

Q-constant band-pass filters $\hat\psi_\lambda$

$Q \approx 16$ for audio

- Wavelet transform: $Wx(t) = \left\{x \star \phi\,(t)\ ,\ x \star \psi_\lambda(t)\right\}_\lambda$

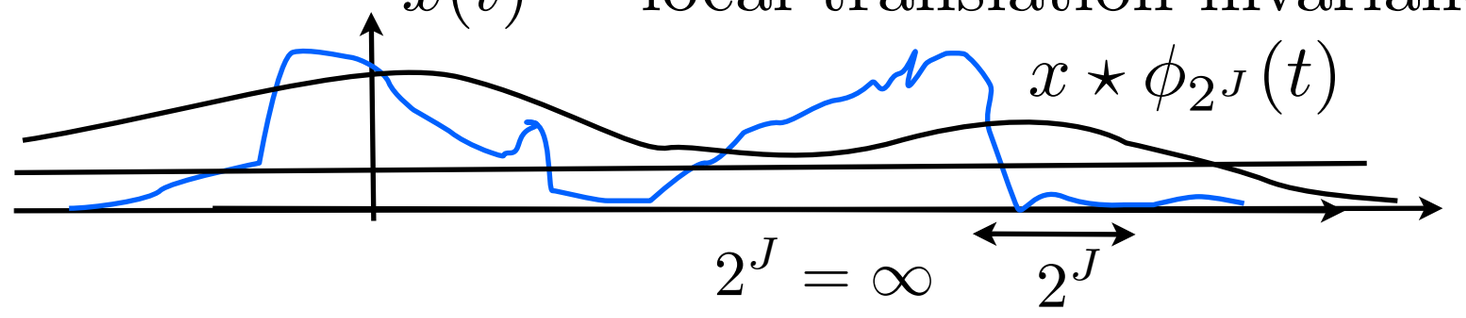
- If $|\phi|^2 + \sum_\lambda |\hat\psi_\lambda|^2 = 1$ then $\|Wx\|^2 = \|x\|^2$.

Wavelet Translation Invariance

First wavelet transform

$$W_1 x = \begin{pmatrix} x \star \phi_{2^J} \\ x \star \psi_{\lambda_1} \end{pmatrix}_{\lambda_1}$$

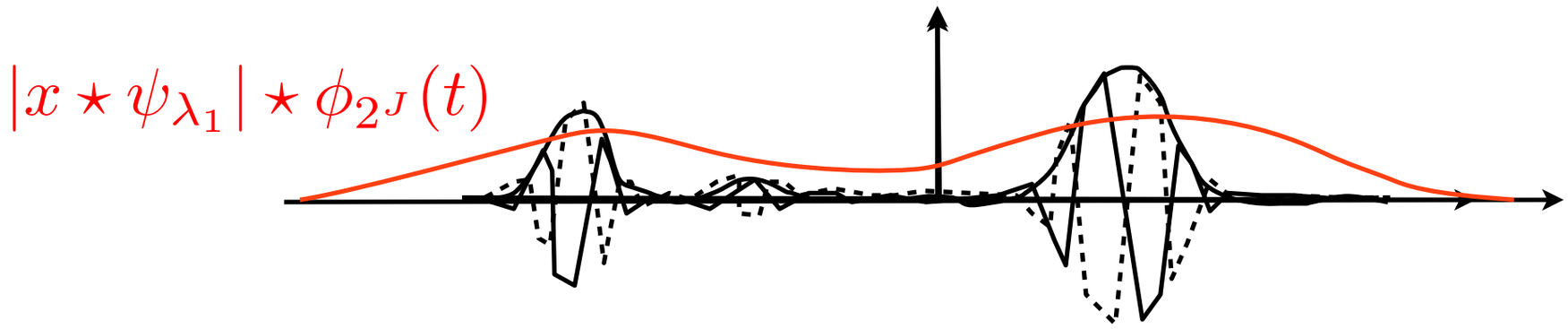
$$|W_1| x = \begin{pmatrix} x \star \phi_{2^J} \\ |x \star \psi_{\lambda_1}| \end{pmatrix}_{\lambda_1}$$



full translation invariance

local translation invariance

Modulus improves invariance: $|x \star \psi_{\lambda_1}| \star \phi_{2^J}(t) = \sqrt{|x \star \psi^a_{\lambda_1}|^2(t) + |x \star \psi^b_{\lambda_1}|^2(t)}$



Second wavelet transform modulus

$$|W_2| \; |x \star \psi_{\lambda_1}| = \begin{pmatrix} |x \star \psi_{\lambda_1}| \star \phi_{2^J}(t) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t)| \end{pmatrix}_{\lambda_2}$$

Scattering Convolution Network

1D

$x(t)$

t

$|W_1|$

$Q_1 = 16$

2D

$|x \star \psi_{\lambda_1}(t)|$

λ_1

t

time average

$2^J = \infty$

Output

$2^J \leq 50ms$

Mel Frequency Spectrum

$|W_2|$

$Q_2 = 1$

$|x \star \psi_{\lambda_1}| \star \phi_{2^J}(t)$

λ_1

t

3D

$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t)|$

λ_1

t

λ_2

time average

$|W_3|$

$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J}(t)$

Output:

Modulation Spectrum

λ_1

t

t

λ_2

Scattering Properties

$$S_J x = \begin{pmatrix} x \star \phi_{2^J} \\ |x \star \psi_{\lambda_1}| \star \phi_{2^J} \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J} \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi_{2^J} \\ \dots \end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, \dots} = \dots |W_3| \, |W_2| \, |W_1| \, x$$

$$\|W_k x\| = \|x\| \quad \Rightarrow \quad \| |W_k x| - |W_k x'| \| \leq \|x - x'\|$$

Theorem: *For appropriate wavelets, a scattering is*

contractive $\|S_J x - S_J y\| \leq \|x - y\|$ ($\mathbf{L^2}$ *stability*)

preserves norms $\|S_J x\| = \|x\|$

translations invariance and deformation stability:

if $D_\tau x(u) = x(u - \tau(u))$ *then*

$$\lim_{J \to \infty} \|S_J D_\tau x - S_J x\| \leq C \, \|\nabla \tau\|_\infty \, \|x\|$$

Audio Model

Excitation $e(t)$ (pitched or random)

Resonator filter $h(t)$

Amplitude modulation $a(t)$

$$x(t) = a(t)\, e \star h(t)$$

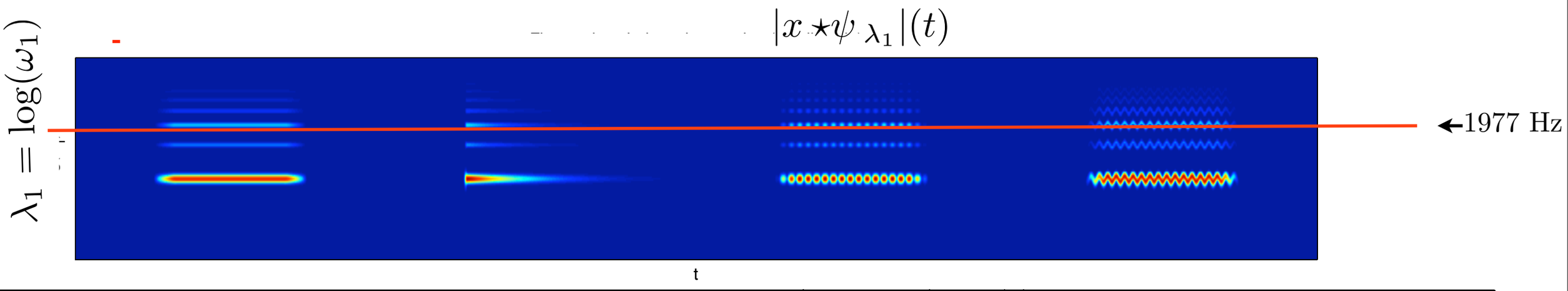
If the excitation is stationary then

$$\frac{||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J}}{|x \star \psi_{\lambda_1}| \star \phi_{2^J}} \approx \frac{|a \star \psi_{\lambda_2}| \star \phi_{2^J}}{a \star \phi_{2^J}}$$

which characterises the amplitude spectrum.

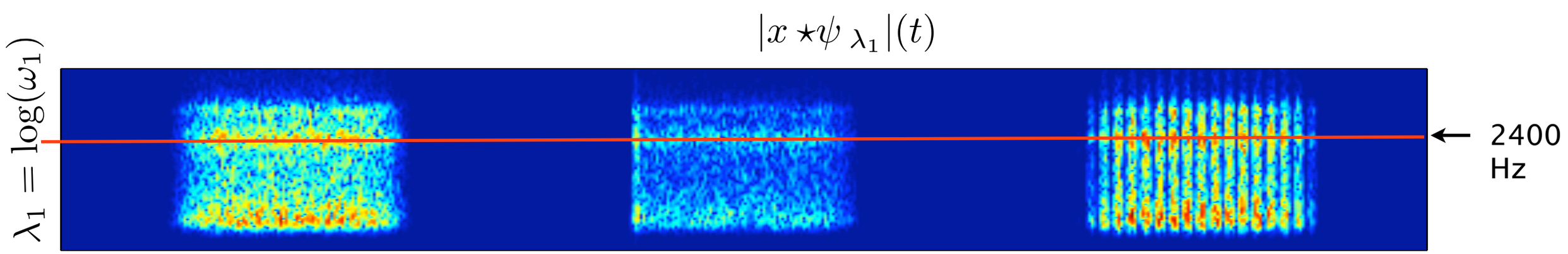
Amplitude Modulation

Harmonic sound: $x(t) = a(t)\, e \star h(t)$ with varying $a(t)$



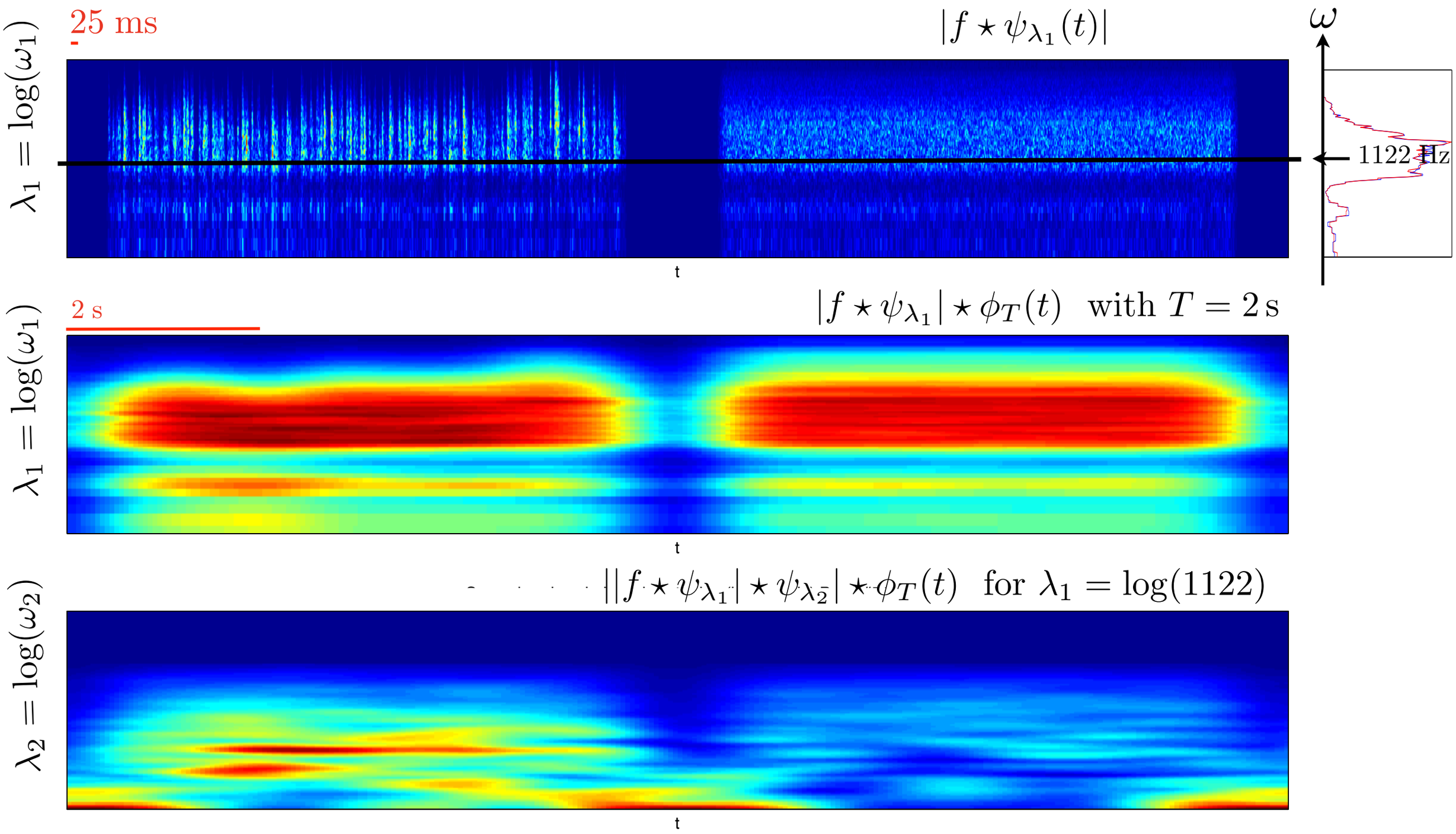
Stochastic Excitation

Random source: $x(t) = a(t)\, e \star h(t)$ with varying $a(t)$



Same Power Spectrum

Random source: $x(t) = a(t)\, e \star h(t)$ with varying $a(t)$



Genre Classification (GTZAN)

- Musical genre classification (jazz, rock, classical, ...)
  10 classes and 30 seconds tracks.
- Each frame is classified using a Gaussian kernel SVM.

Feature Set	Error (%)
Delta-MFCCs	18.0
Time Scat., order 1	19.1
Time Scat., order 2	10.7
Time Scat., order 3	10.6
Time-frequency Scat., order 2	9.4

$T = 740$ ms

- Difficult to analyse the sources of inefficiencies

Time Invariant Scattering Moments

The scattering transform of a stationary process $X(t)$

$$S_J X = \begin{pmatrix} X \star \phi_{2^J} \\ |X \star \psi_{\lambda_1}| \star \phi_{2^J} \\ ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J} \\ |||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi_{2^J} \\ \dots \end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, \dots}$$

converges to moments if X is ergodic when 2^J increases

$$\mathbb{E}(SX) = \begin{pmatrix} \mathbb{E}(X) \\ \mathbb{E}(|X \star \psi_{\lambda_1}|) \\ \mathbb{E}(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ \mathbb{E}(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) \\ \dots \end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, \dots}$$

- Does $\mathbb{E}(SX)$ approximates well enough the distribution of X ?

Inverse Scattering Transform

Joan Bruna

- Reconstruct $x(t)$ from:

N samples

- 1st order coefficients: $|x \star \psi_{\lambda_1}| \star \phi_{2^J}(t), \ \forall \lambda_1$

$$\text{for } 2^J = \infty\text{: } \int |x \star \psi_{\lambda_1}(u)|\, du$$

$Q_1 \log_2 N$ coeffs

- 2nd order coefficients: $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J}(t), \ \forall \lambda_1, \lambda_2$

$$\text{for } 2^J = \infty\text{: } \int ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(u)|\, du$$

$Q_1\, Q_2 \log_2^2 N/2$ coeffs

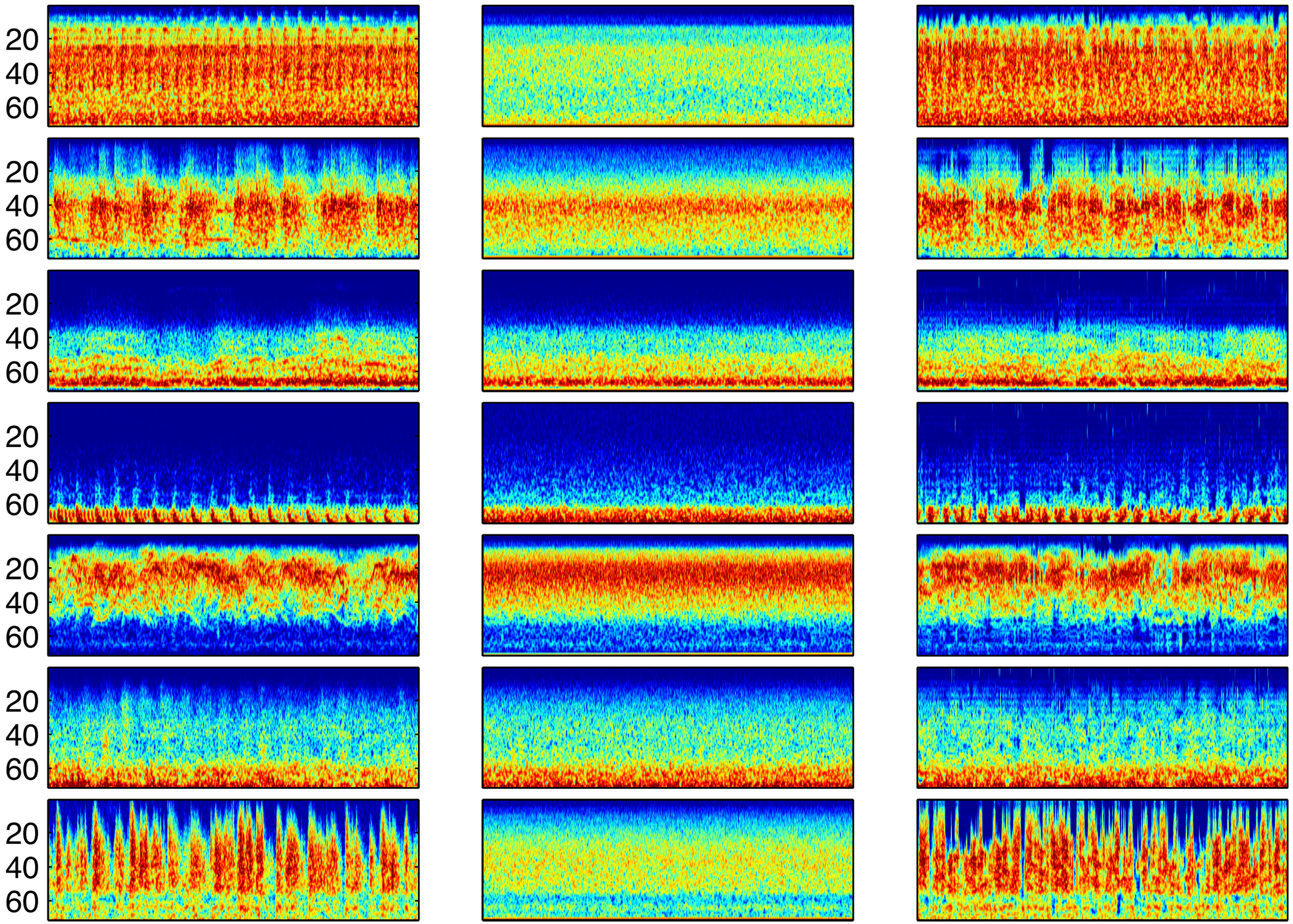
Stochastic of Audio Textures

Joan Bruna

original
[McDermott & Simoncelli'11]

1st order scattering
$K = 50$

1st+2nd order scattering
$K = 500$

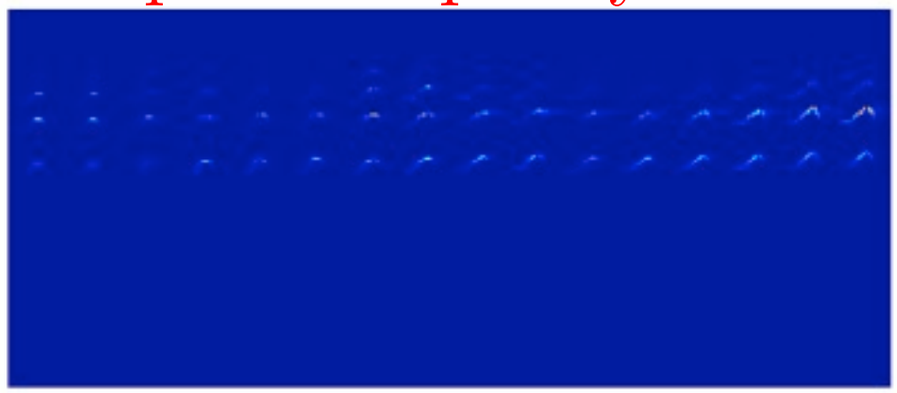


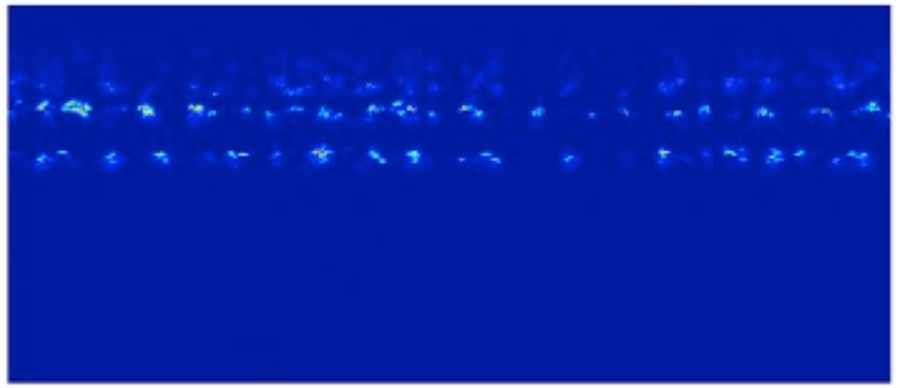
Harmonic Sounds

V. Lostanlen

Need to express frequency channel interactions: time-frequency image

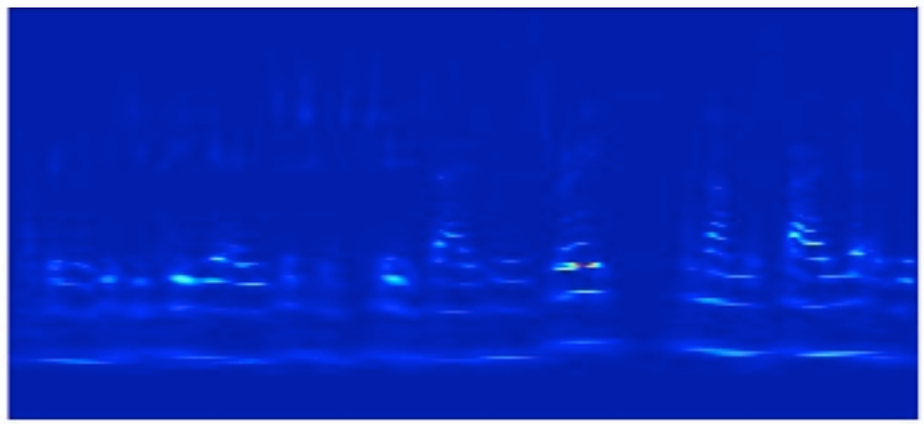
Bird

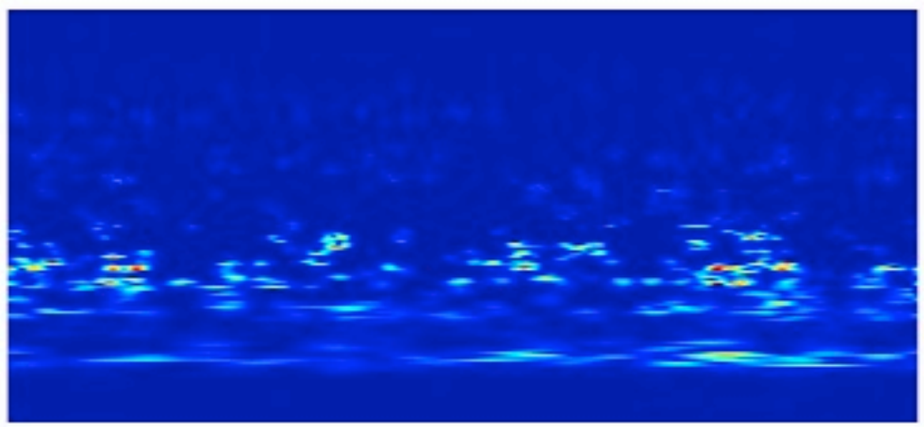




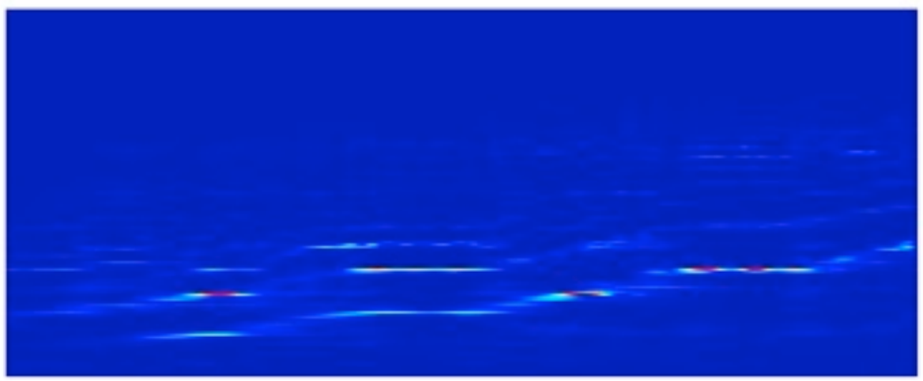
$2^J = \infty$

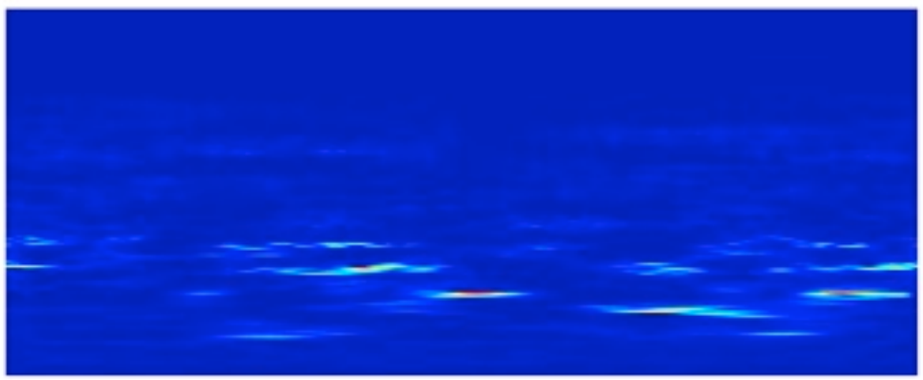
Speech



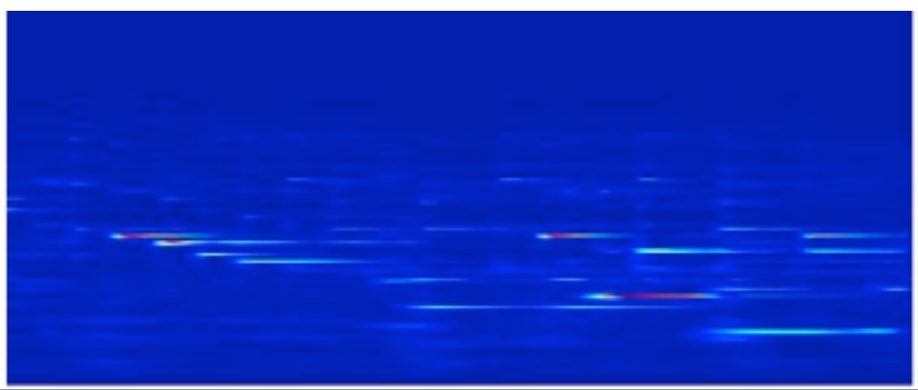


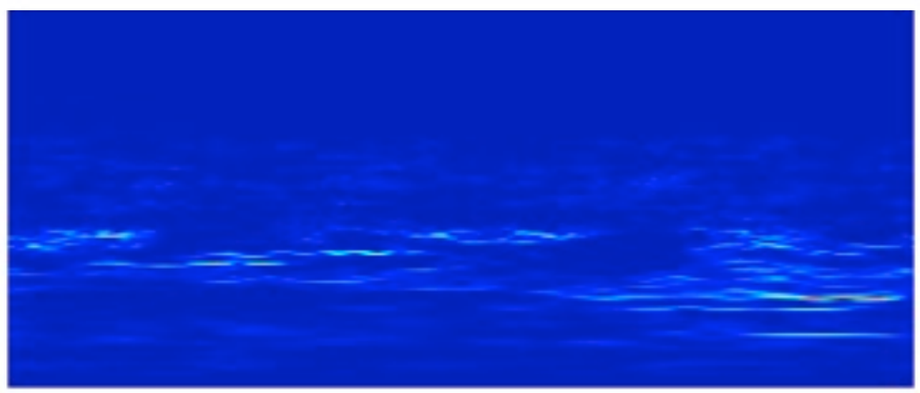
Cello





Piano



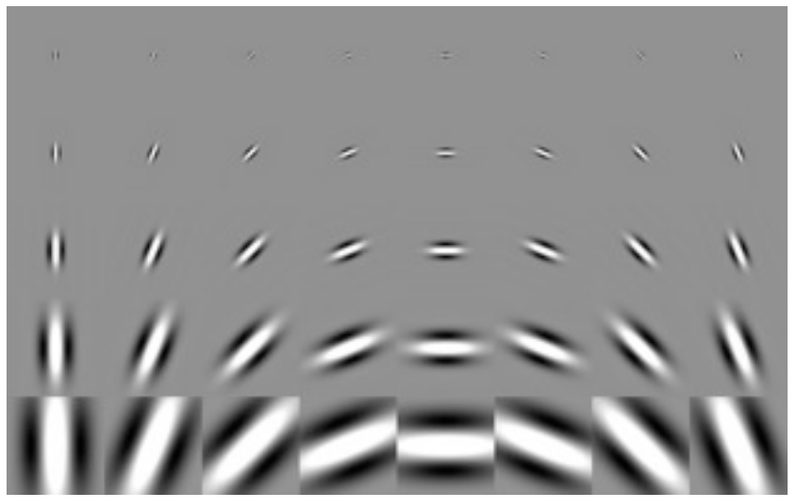


Wavelets for Images

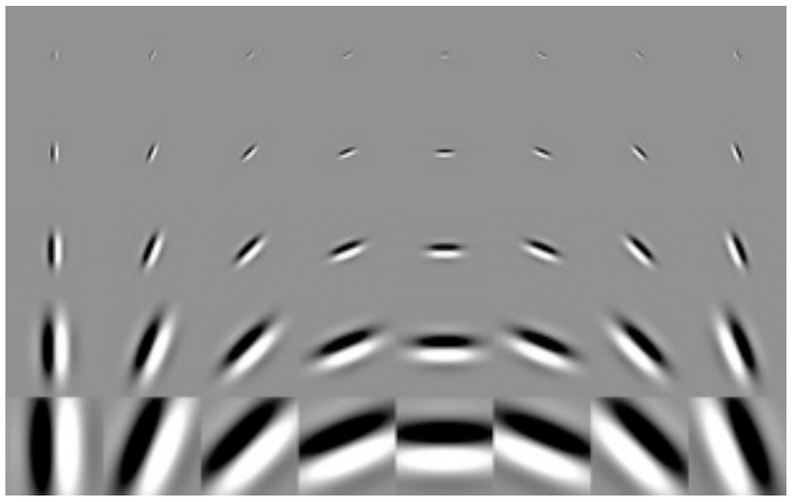
- Complex wavelet: $\psi(t) = g(t) \exp i\xi t \quad , \quad t = (t_1, t_2)$

  rotated and dilated: $\psi_\lambda(t) = 2^{-j}\, \psi(2^{-j} r_\theta t)$ with $\lambda = (2^j, \theta)$

real parts



imaginary parts



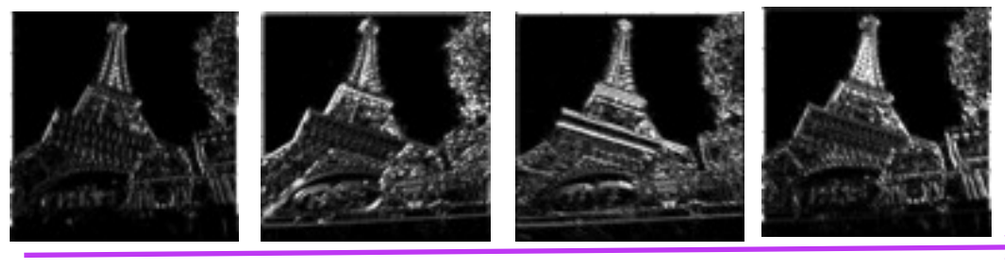
- Wavelet transform: $Wx = \begin{pmatrix} x \star \phi_{2^J}(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{\lambda \leq 2^J}$

Scattering Transform in 2D

$|W_1|$

$|x \star \psi_{2^1,\theta}|$

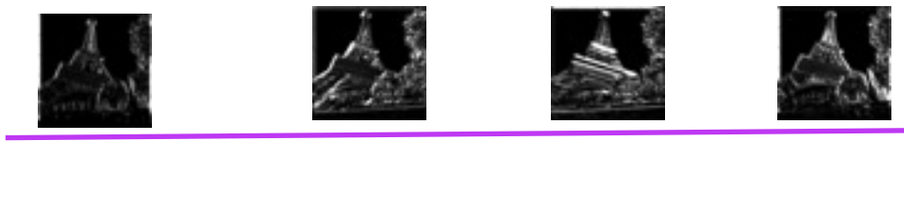
$|W_2|$



θ

3D wavelet convolutions
2D space and angle

$|x \star \psi_{2^2,\theta}|$



θ

x

$|x \star \psi_{2^3,\theta}|$

θ

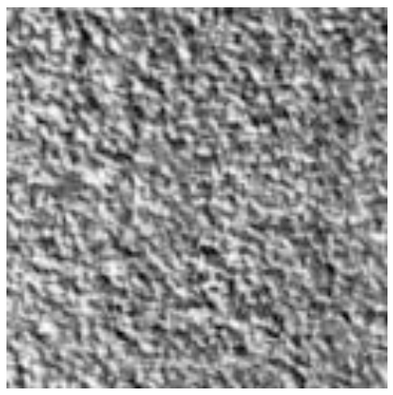
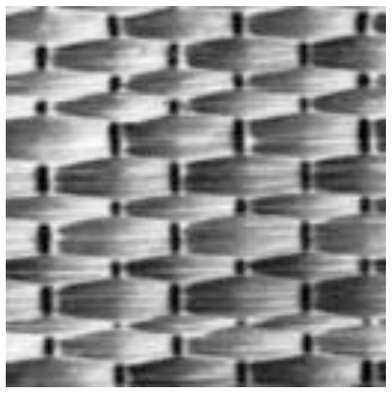
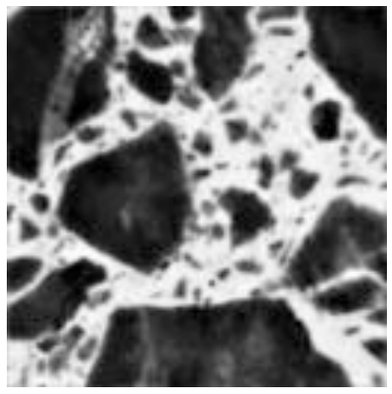
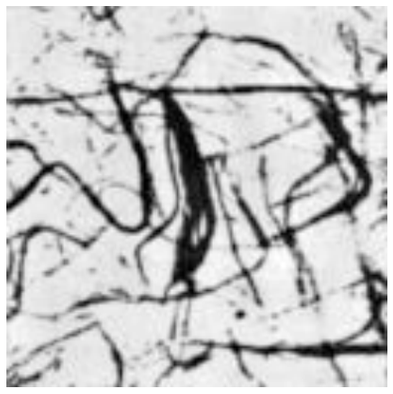
$x \star \phi_J$

θ

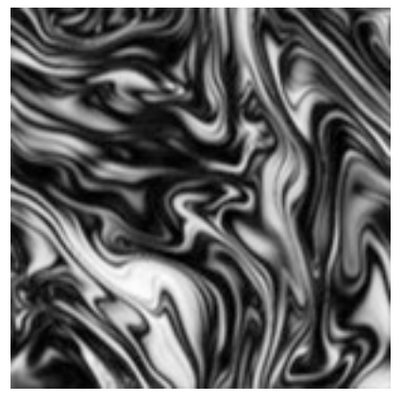
Ergodic Texture Reconstructions

Joan Bruna

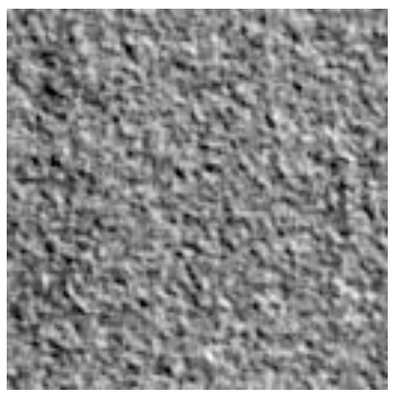
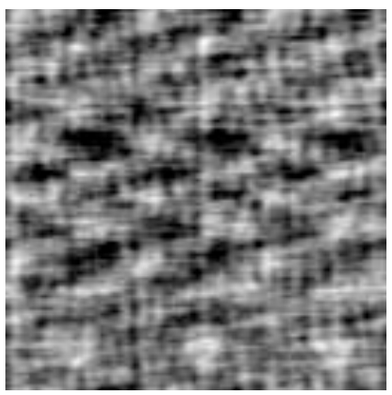
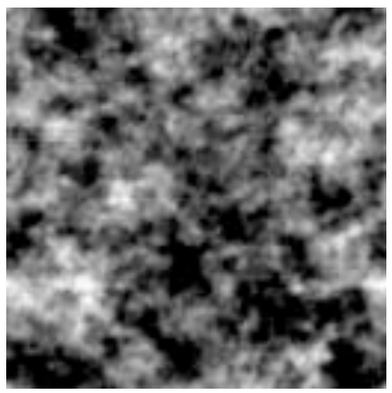
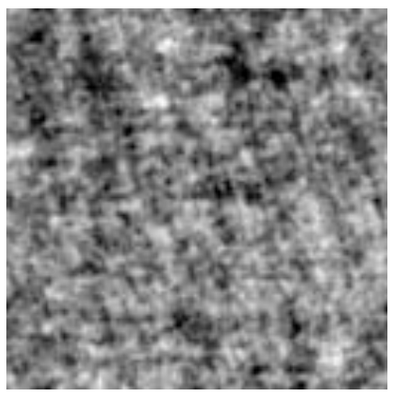
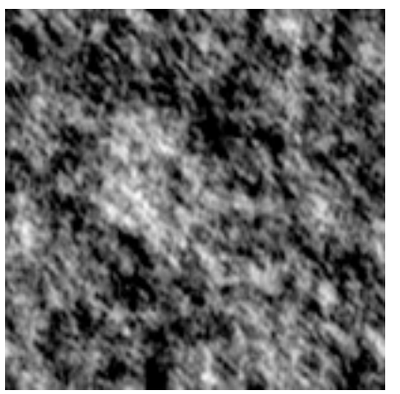
Original Textures

2D Turbulence

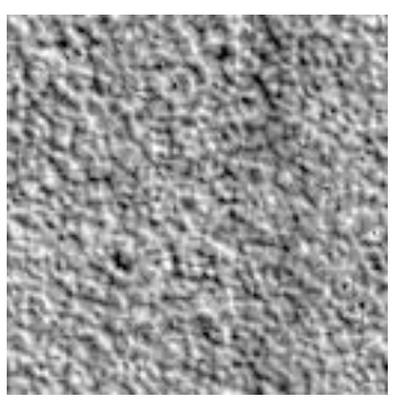
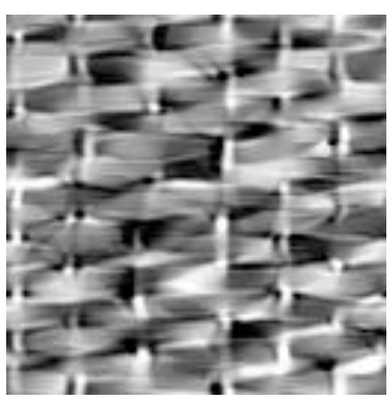


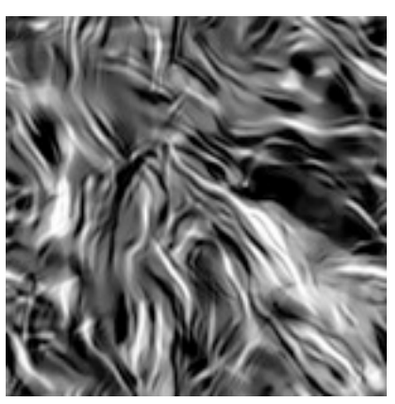
Gaussian process model with same second order moments

For $2^J = N$: $O(\log N^2)$ scattering moments:

$$\|x \star \psi_{\lambda_1}\|_1 \approx \mathbb{E}(|x \star \psi_{\lambda_1}|) \quad , \quad \||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1 \approx \mathbb{E}(||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|)$$

Cortical Transform

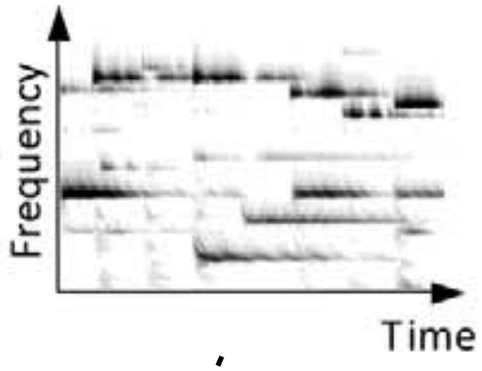
K. Patil, D. Pressnitzer, S. Shamma, M. Elhilali

1D

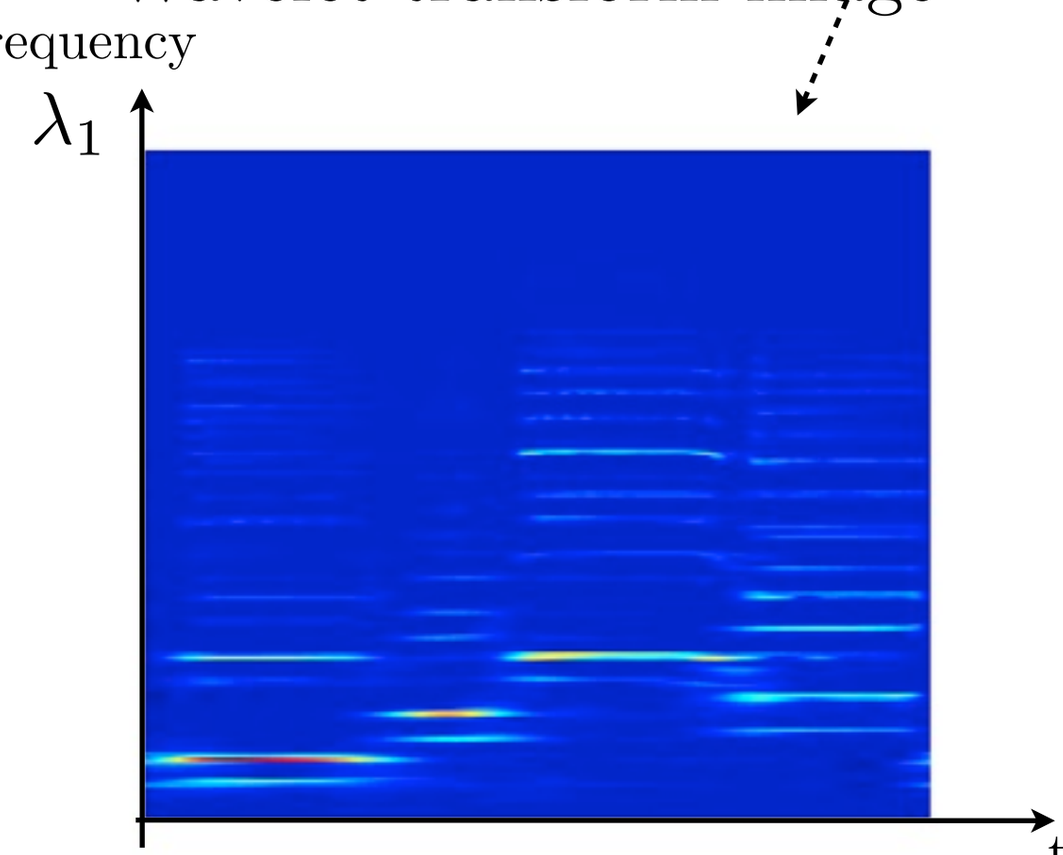
Cochlear/
Midbrain
Processing

$|W_1|$

2D



Wavelet transform image

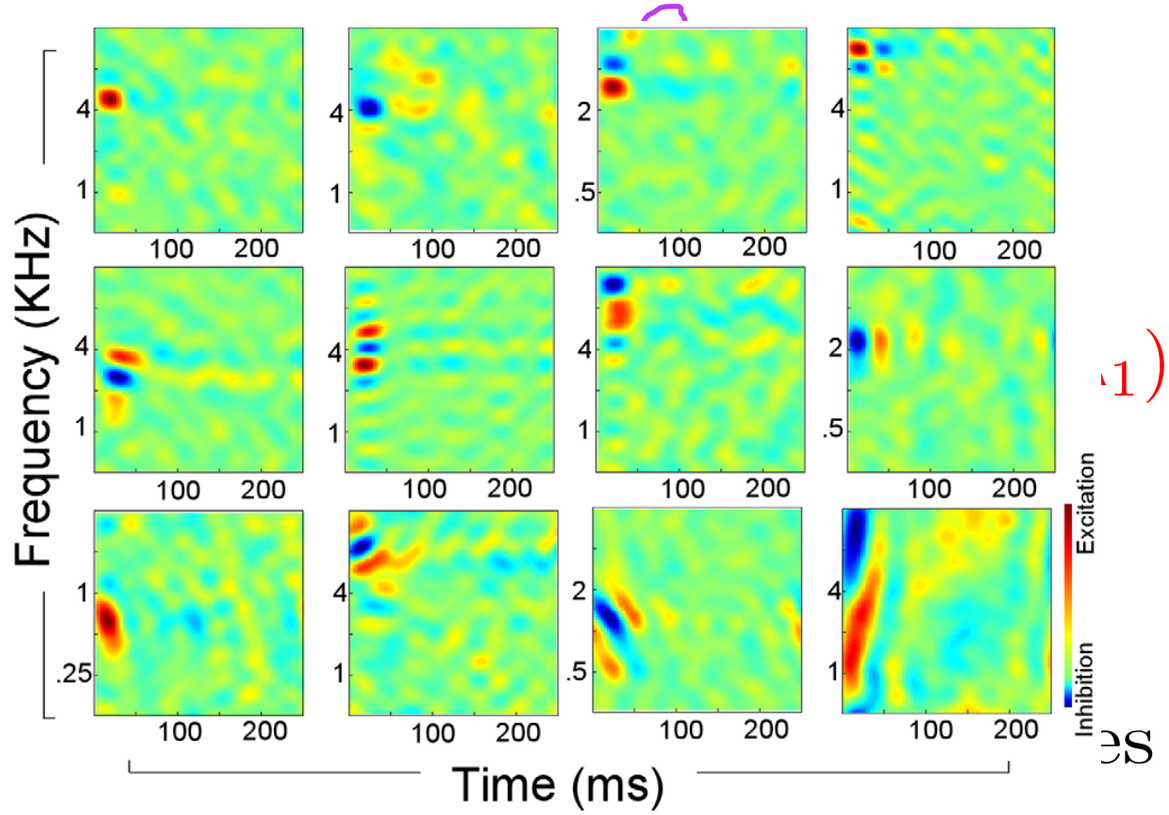


$\star$

Separable Time-Frequeny 2D wavelets

Neuron Receptive Fields in A1

$\psi_{\lambda_t}(t)$

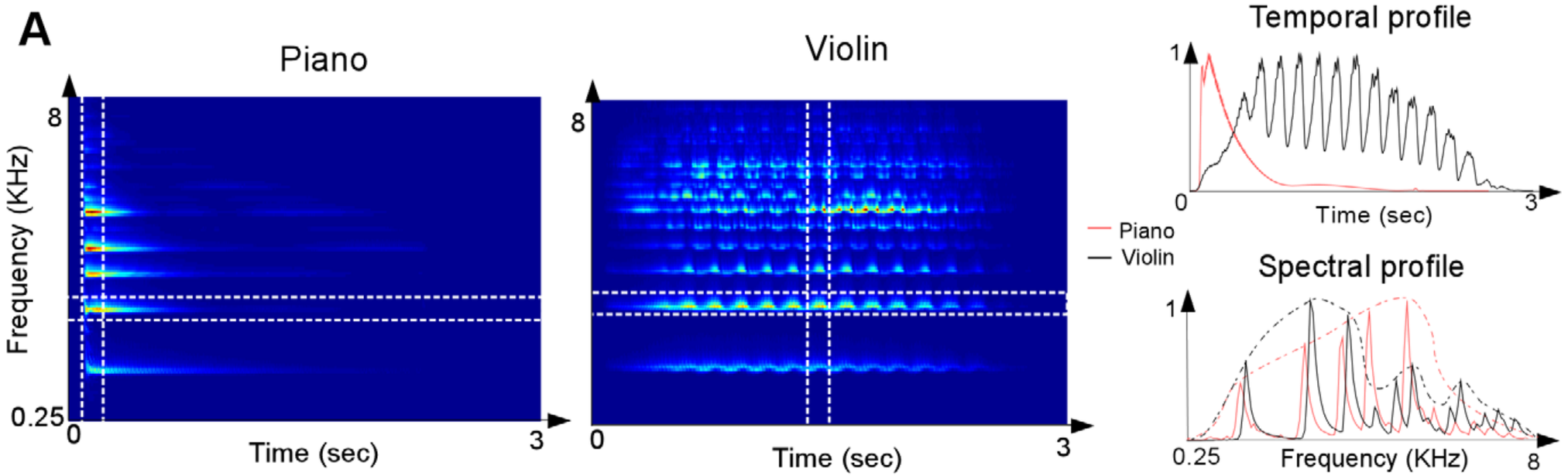


$,1)$

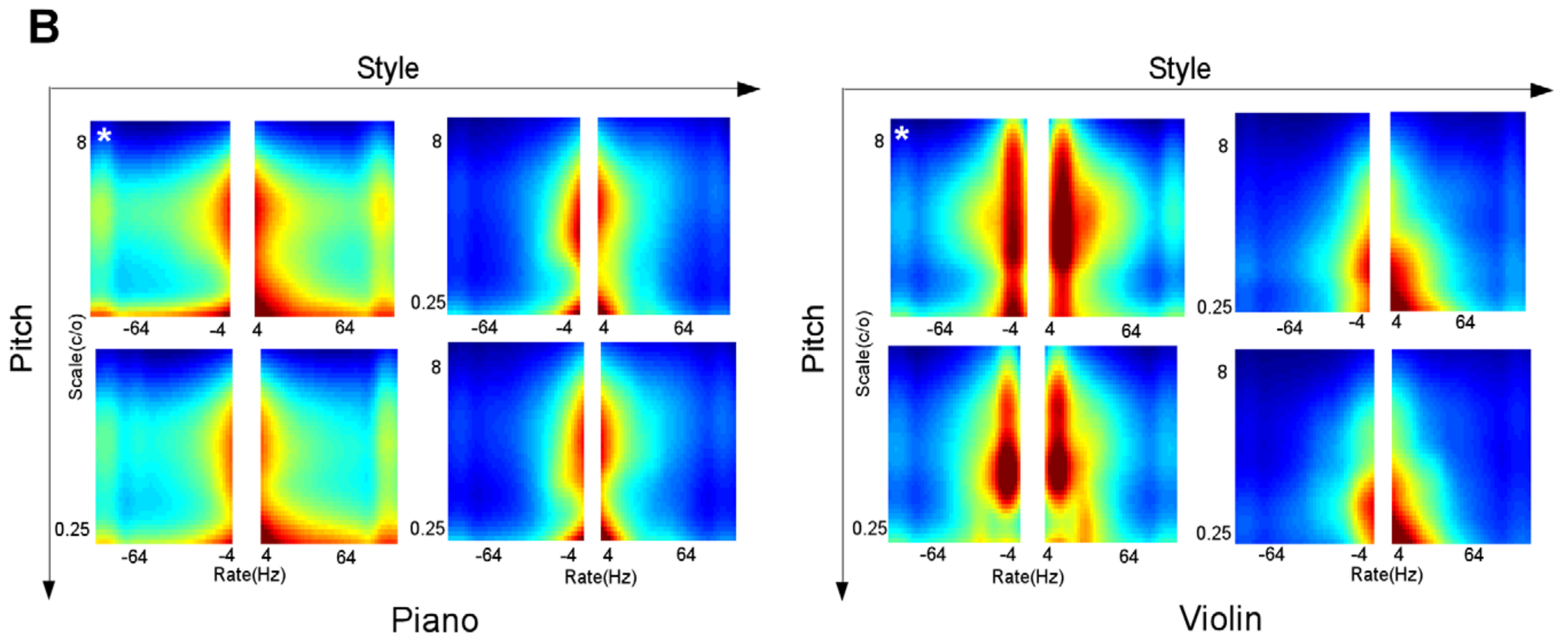
es

Timber of Musical Instruments

2D Scalogram



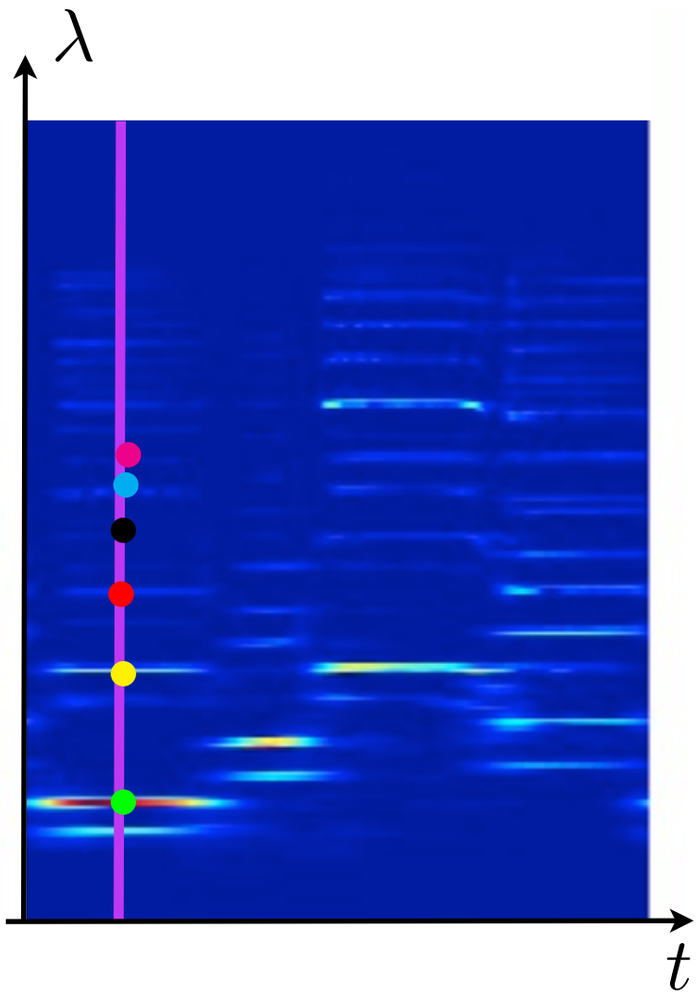
Collapsing 4D into 2D images by averaging time and frequencies

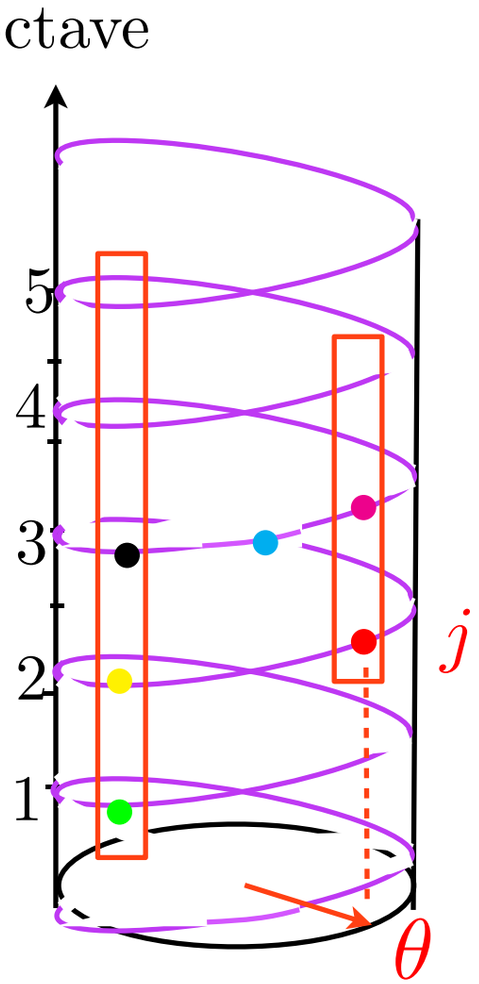


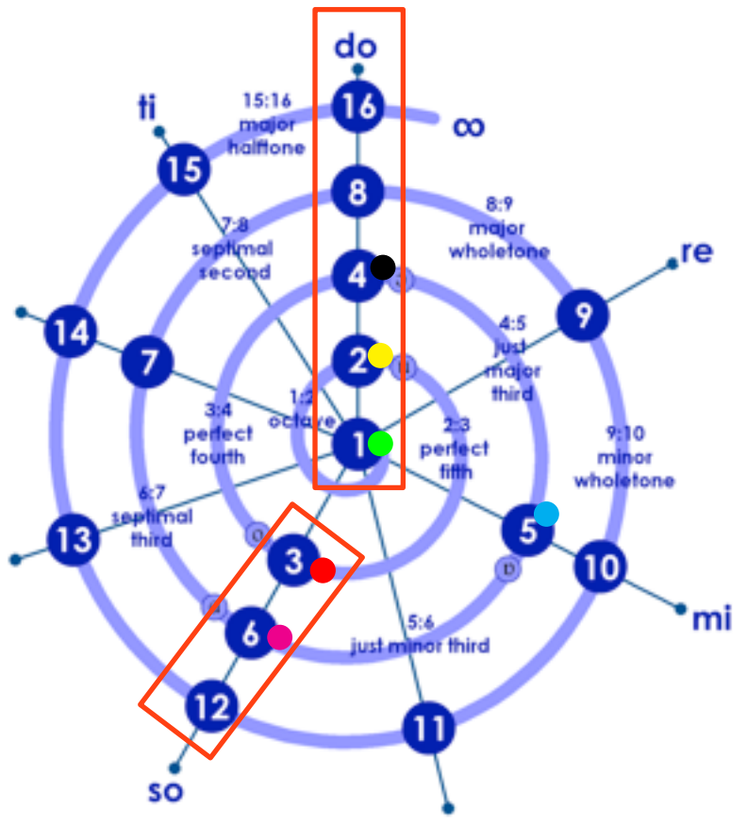
- Good instrument classification performances with an SVM

Harmonic Spiral

- Problem: the wavelet transform of harmonics is not sparse





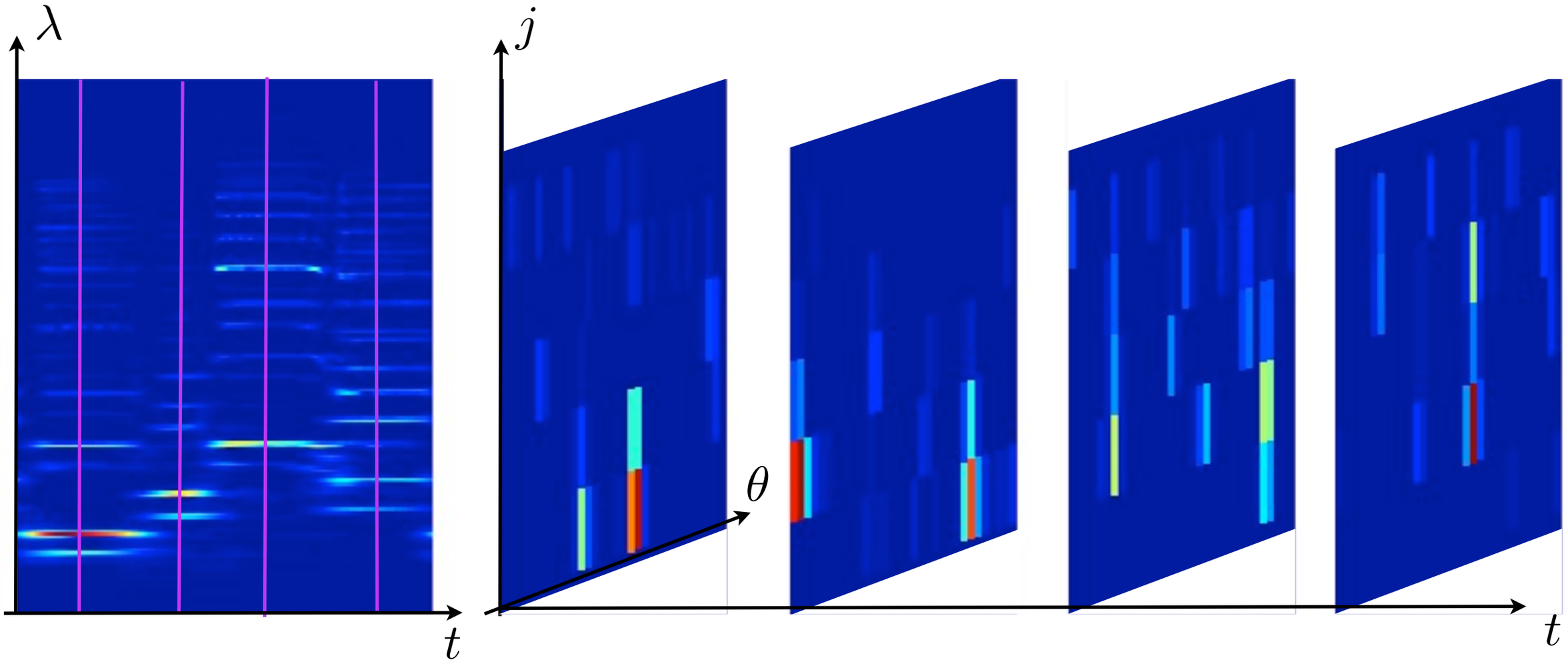


- Alignment of harmonics in two main groups.

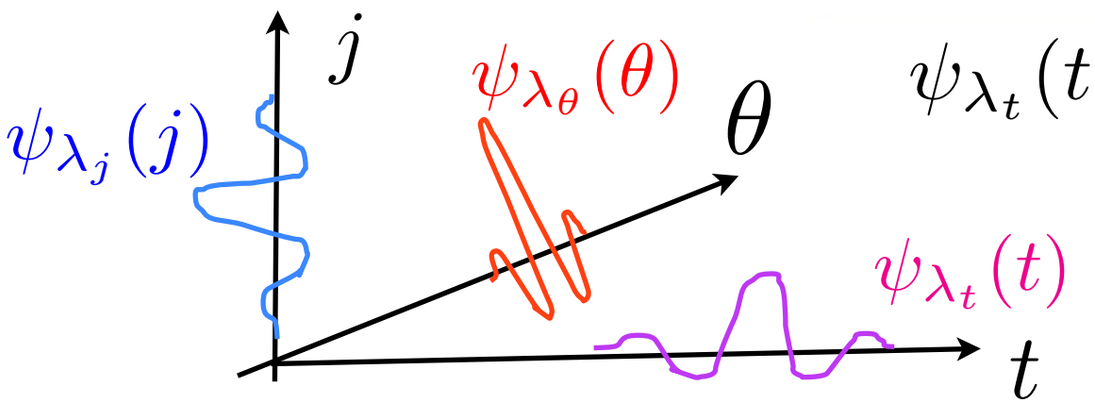
  More regular variations along (θ, j) than λ

Spiral Scattering

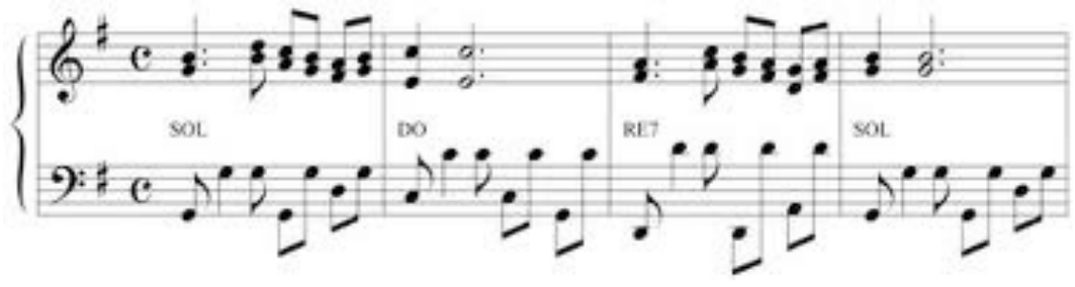
V. Lostanlen



3D separable Spiral wavelet transform W_2

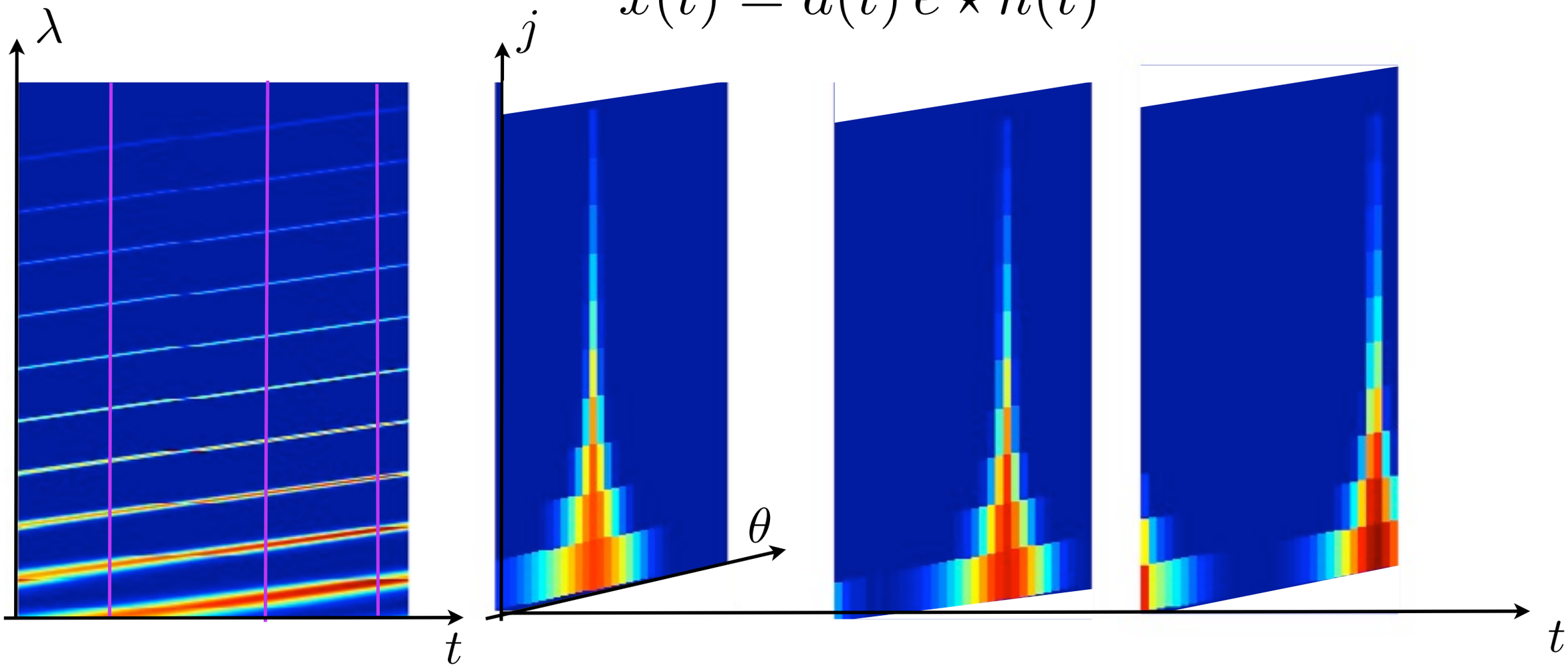


$$\psi_{\lambda_t}(t)\,\psi_{\lambda_\theta}(\theta)\,\psi_{\lambda_j}(j)$$

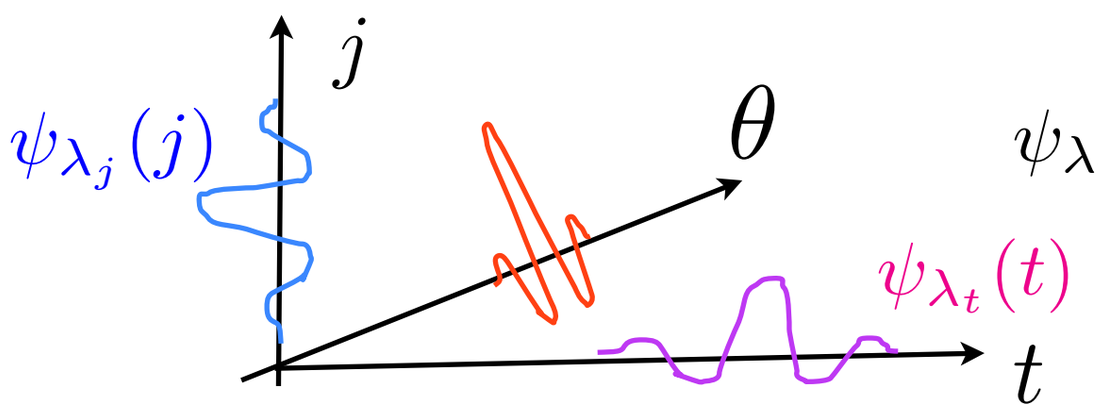


Shepard-Risset Glissando

$$x(t) = a(t)\, e \star h(t)$$



3D separable Spiral wavelet transform W_2



$$\psi_{\lambda_j}(j)$$

$$\psi_{\lambda_t}(t)$$

$$\psi_{\lambda_t}(t)\, \psi_{\lambda_\theta}(\theta)\, \psi_{\lambda_j}(j)$$

Separate Pitch, Amplitude, Envelop-

V.Lostanlen

- Second order spiral scattering in 5D:

$$Sx(t, \lambda_1, \lambda_t, \lambda_\theta, \lambda_j) = ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_t} \psi_{\lambda_\theta} \psi_{\lambda_j}| \star \phi_{2^J}(t)$$

- Scale separation in $x(t) = a(t)\, e \star h(t)$

$$\frac{\partial S(t, \lambda_1, \lambda_t, \lambda_\theta, \lambda_j)}{\partial \lambda_t} \quad : \quad \text{depends on the moduluation amplitude}$$

$$\frac{\partial S(t, \lambda_1, \lambda_t, \lambda_\theta, \lambda_j)}{\partial \lambda_\theta} \quad : \quad \text{depends on the pitch variations}$$

$$\frac{\partial S(t, \lambda_1, \lambda_t, \lambda_\theta, \lambda_j)}{\partial \lambda_j} \quad : \quad \text{depends on the frequency envelop } \hat{h}(\omega)$$

Remarks and Questions

- Difficult to study timber through classification experiments only
  Analysis-Synthesis is a powerful complement

- What relations between image and audio perception ?
  - Auditory scene analysis *(Bregman)*
  - Can we translate audio in images and reverse ?

- Can we analyse timbre from a geometrical point of view ?
  - Search for regularity transformed into sparsity with wavelets

- Are such tools enough to access to large time scale structures ?
- Do we need to learn Deep Network Filters ?