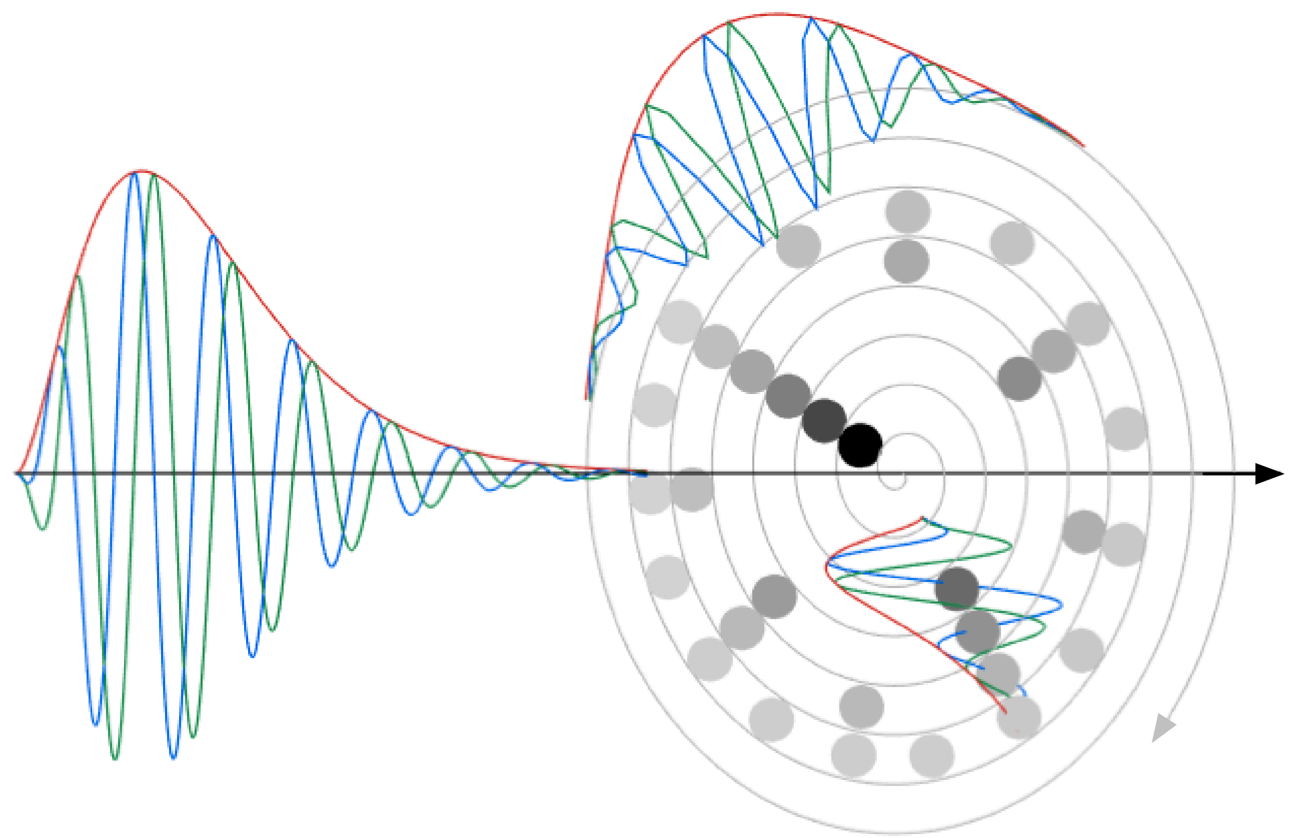
SPIRAL SCATTERING OF NONSTATIONARY SOUNDS

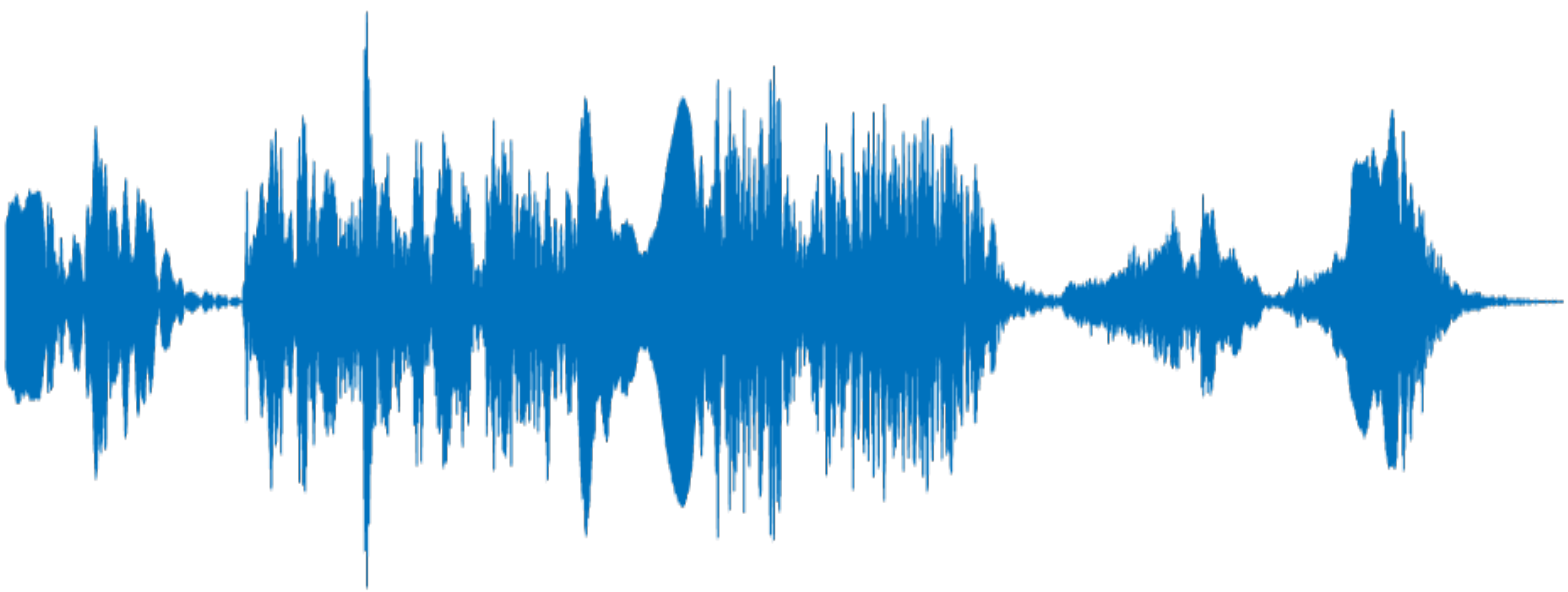


Vincent Lostanlen, Stéphane Mallat
École normale supérieure de Paris

2015-06-19, Marseille, FR. Ce travail est financé par la bourse ERC InvariantClass.

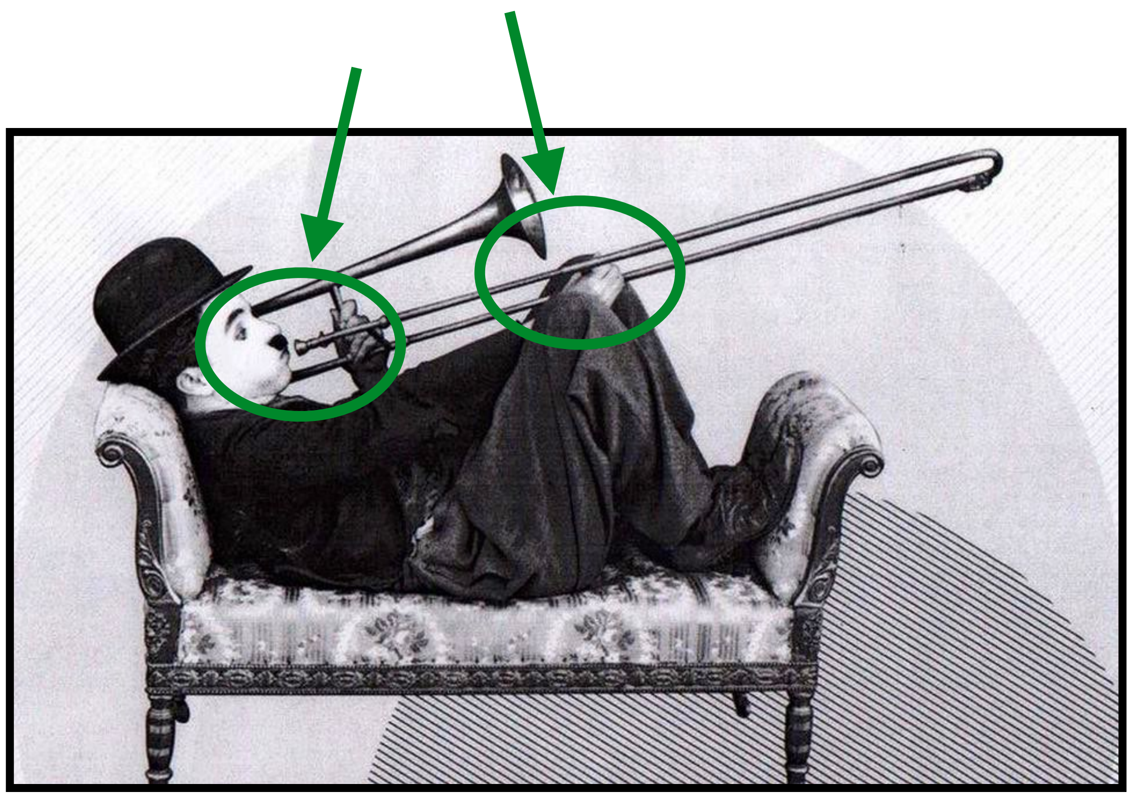
- musical instrument recognition
- query by example
- automated transcription

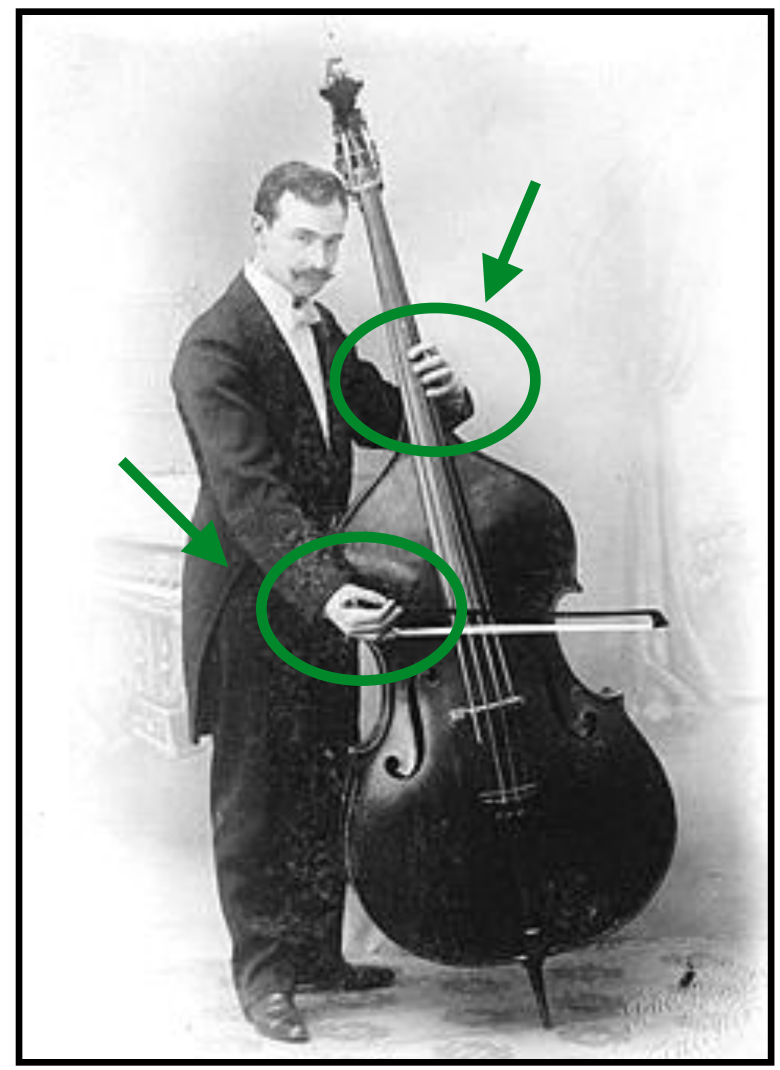
rely on the characterization of musical notes.



Natural regularity along time

The physics of player-instrument interaction brings **regularity** in gestures:

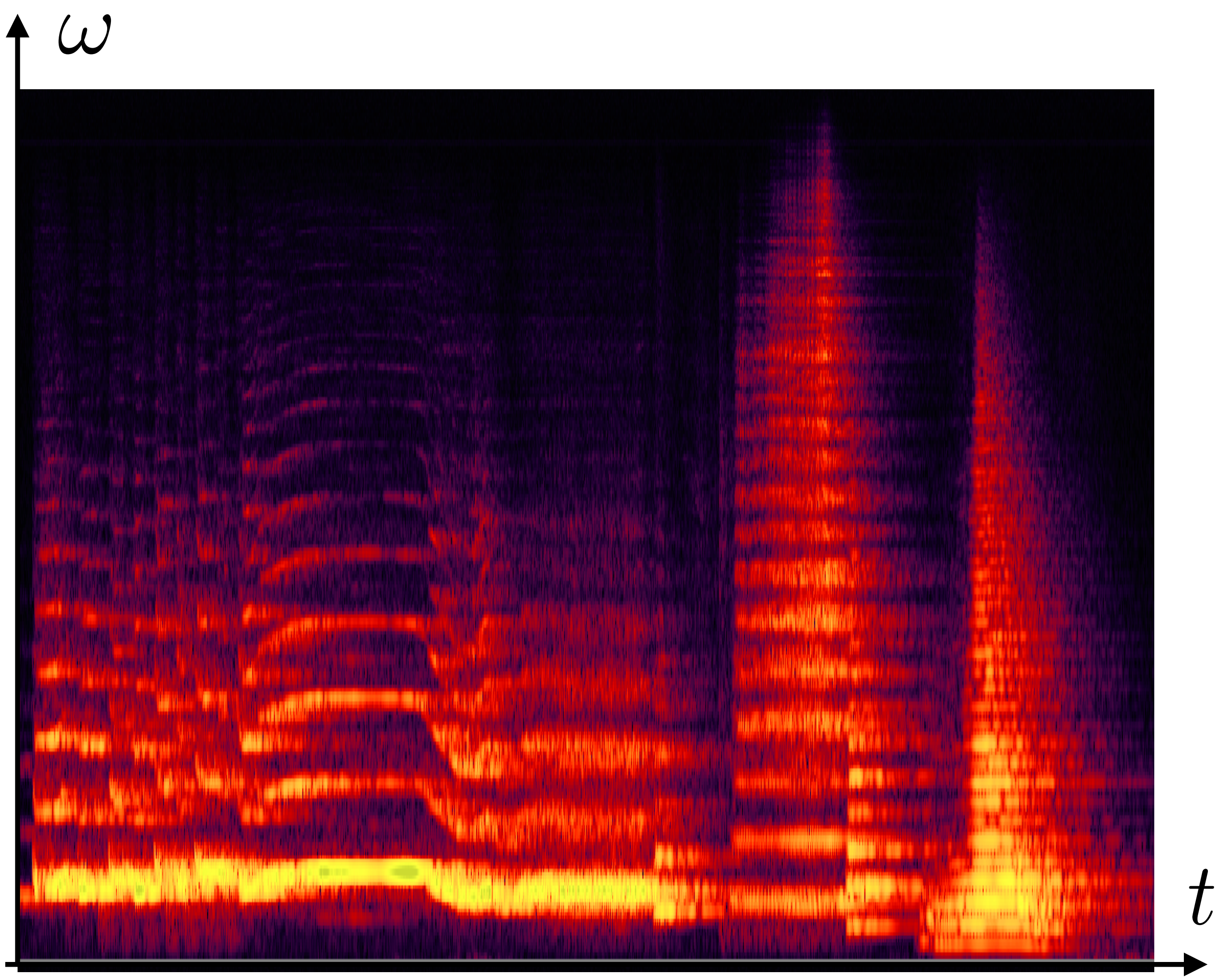




Goal

The TFR reveals short-term regularity (~50 ms).

Goal: to build a meaningful decomposition such that musical notes are regular (~500 ms).



Template-based methods

- Hard constraints, e.g. Markovian.
  - see Kereliuk & Depalle (2008) on partial tracking.
  - useful for fine-grain audio effects...

    ... but exposed to low-level detection errors.

- Loose constraints, e.g. Bayesian.
  - see Fuentes et al. (2013) on harmonic PLCA.
  - encompasses a broad range of meaningful priors...

    ... but timbral regularity remains challenging.

TEMPLATE-FREE APPROACHES

- In fact, most tasks do not require rigid templates.

- We advocate a **progressive decomposition** instead.

- ConvNets perform a data-driven decomposition, but they need a large annotated training set.

- We design a nonlinear **scattering** transform with
  - regularity
  - time-frequency localization
  - sparsity

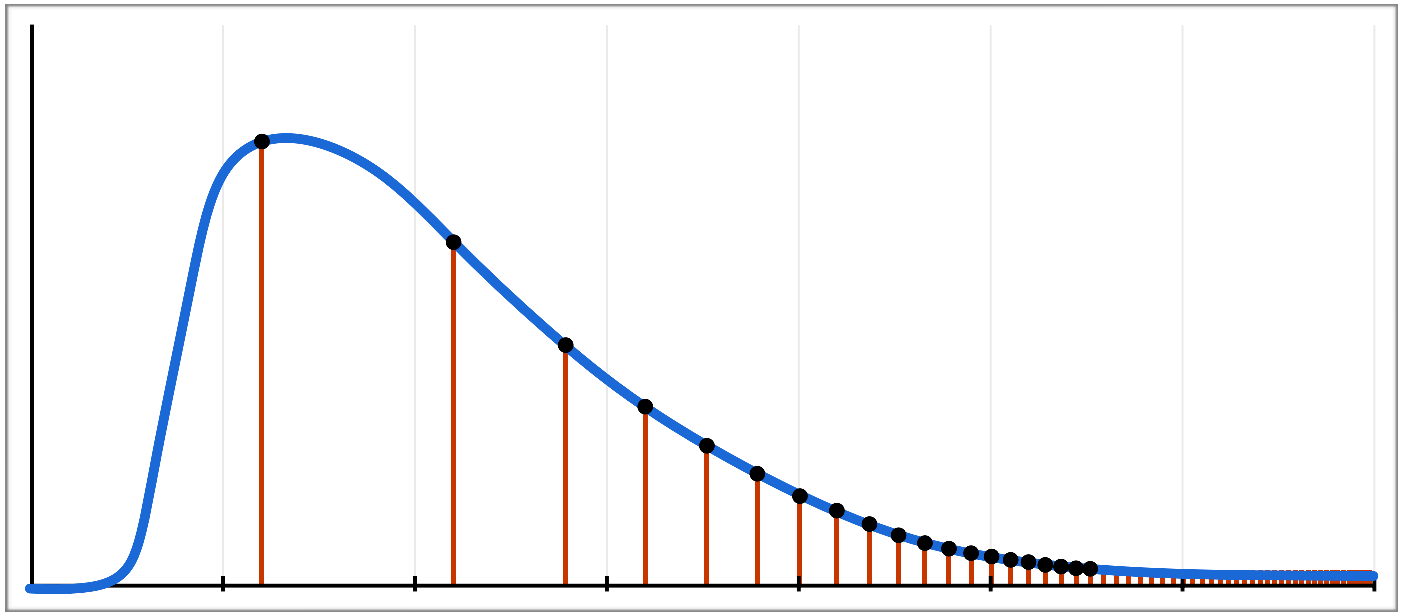
  in mind.

1. Nonstationary source-filter model

2. Leveraging harmonicity

3. Applications to classification and reconstruction

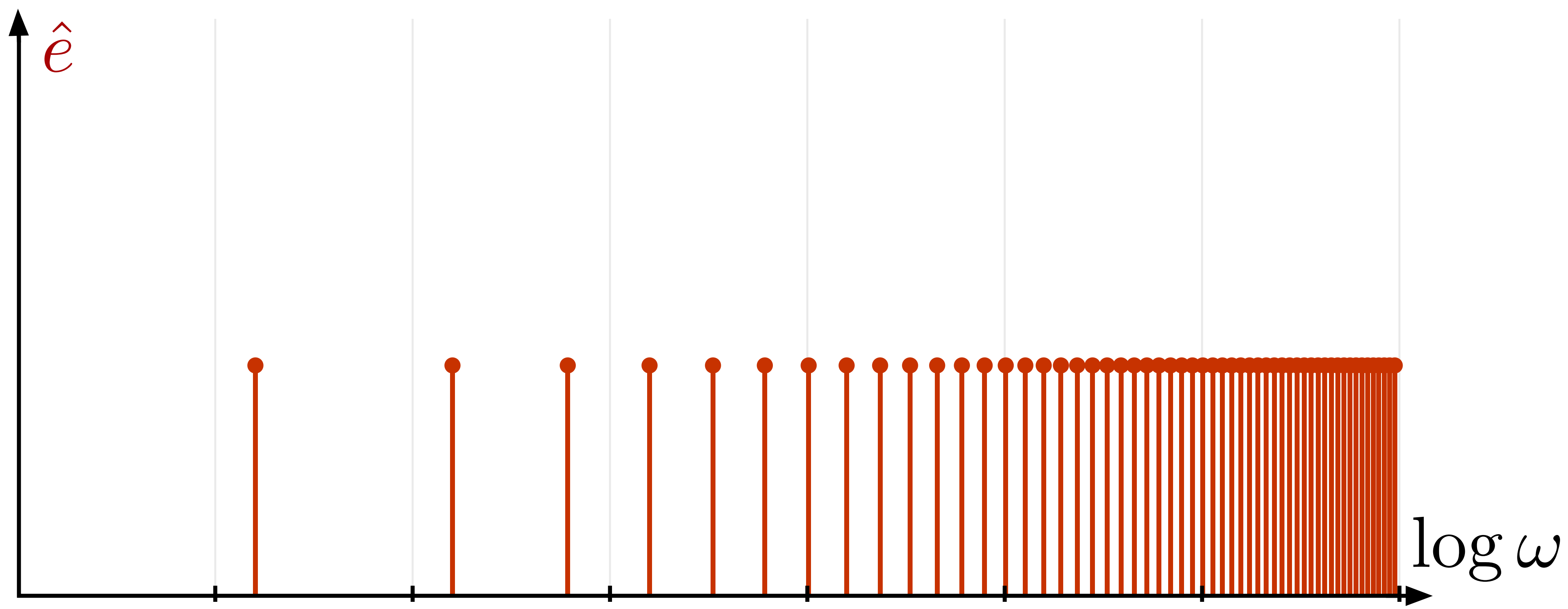
1. Nonstationary source-filter model



Harmonic source

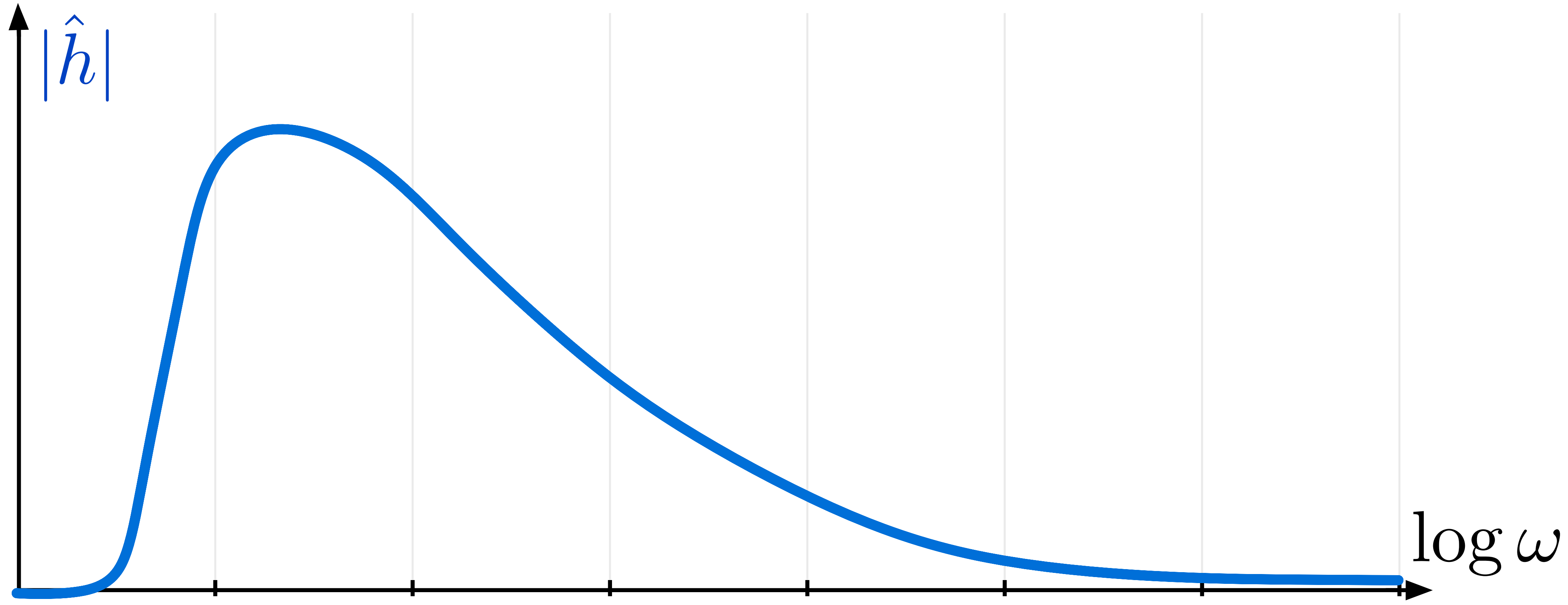
We define the harmonic source as a Dirac pulse train.

$$e(t) = \frac{1}{2\pi} \sum_{n=-\infty}^{+\infty} \delta(t-n) = \sum_{k=-\infty}^{+\infty} \exp(2\pi \mathrm{i} k t)$$



FILTER

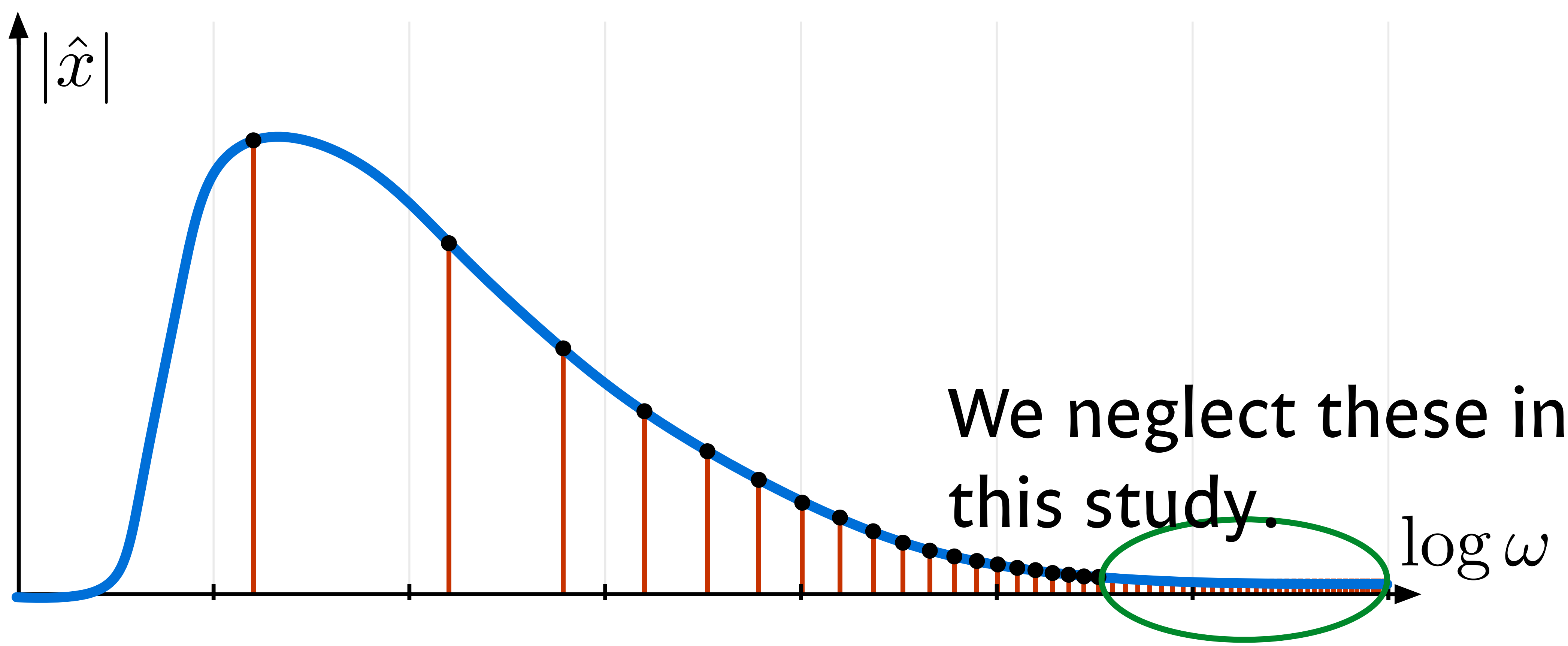
We define the filter as a regular spectral envelope.



Stationary source-filter model

The stationary source-filter model is

$$x(t) = [e * h](t) \text{ i.e. } \hat{x}(\omega) = [\hat{e} \times \hat{h}](\omega)$$



DEFORMATIONS

Let $\theta(t) \in \mathcal{C}^3$ be a time warp function.

$\dot{\theta}(t) > 0$ is the fundamental frequency of

$$e_\theta(t) = (e \circ \theta)(t).$$

$\dot{\nu}(t) > 0$ is the position of the formant (spectral peak)

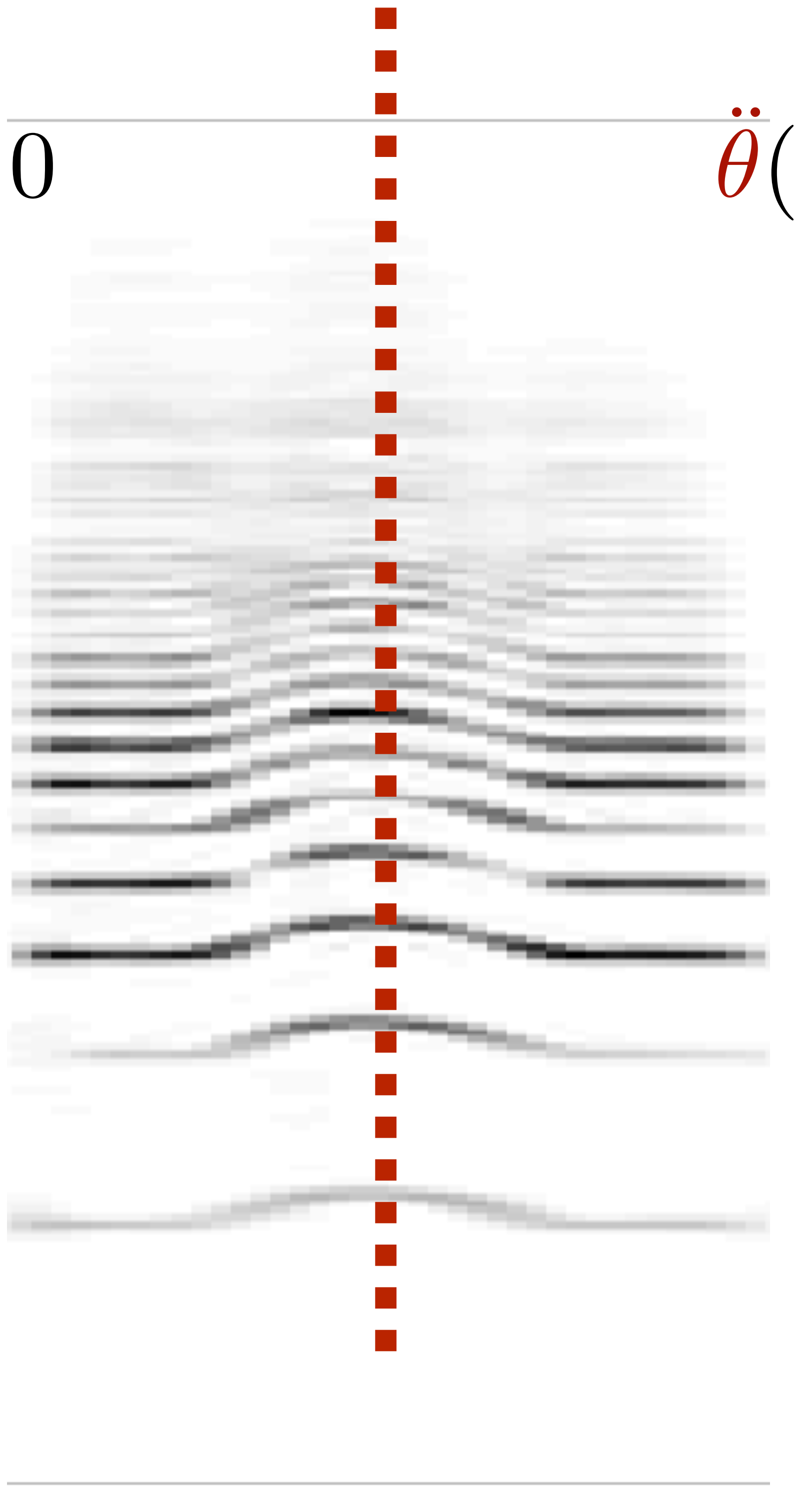
$$h_\nu(t) = (h \circ \nu)(t).$$

The nonstationary source-filter model is defined as

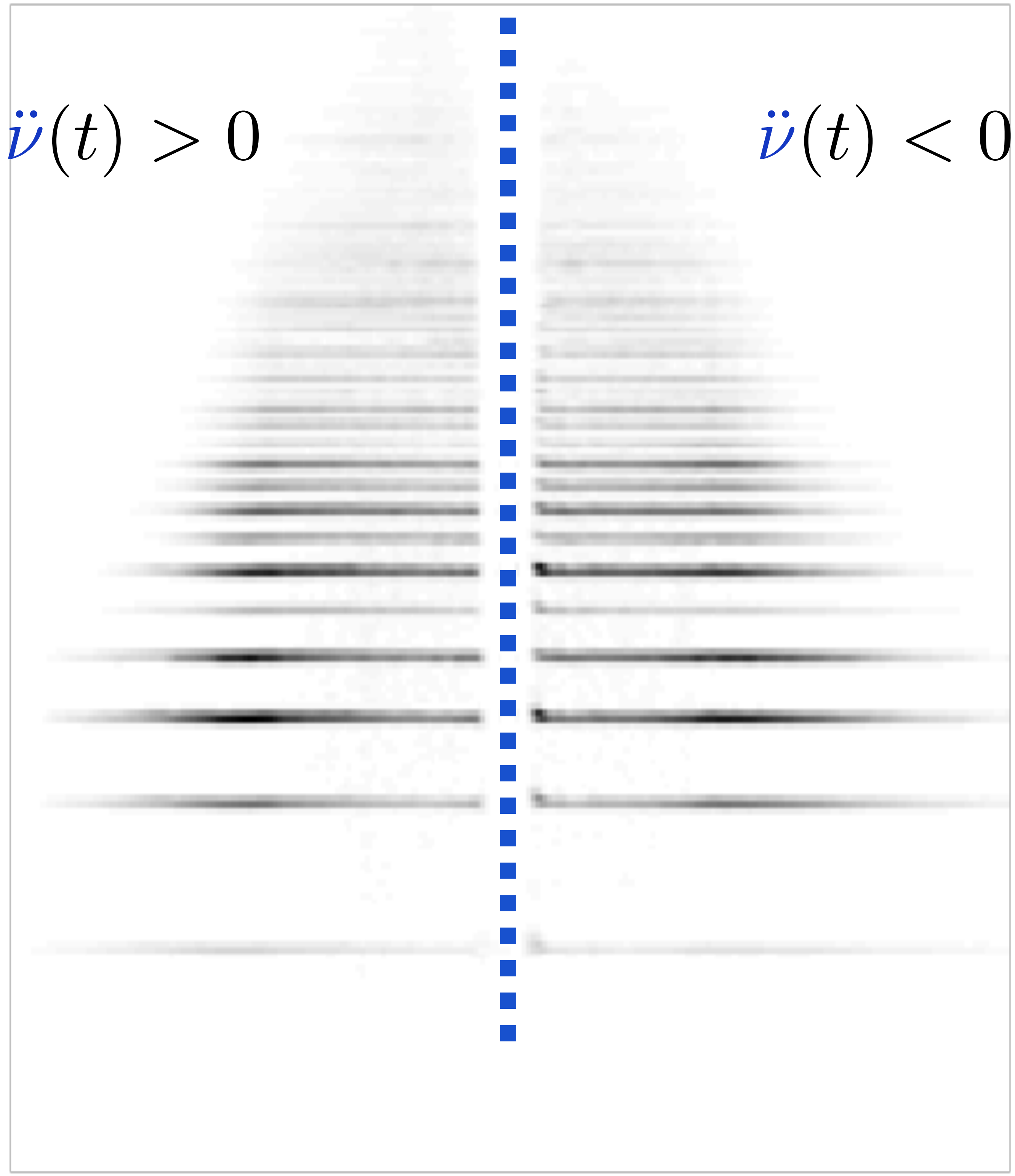
$$x_{\theta,\nu}(t) = [e_\theta * h_\nu](t).$$

EXAMPLE 1 : TROMBONE GLISSANDO

$\ddot{\theta}(t) > 0$ $\ddot{\theta}(t) < 0$



Example 2 : trombone crescendo



A TIME-FREQUENCY PERSPECTIVE

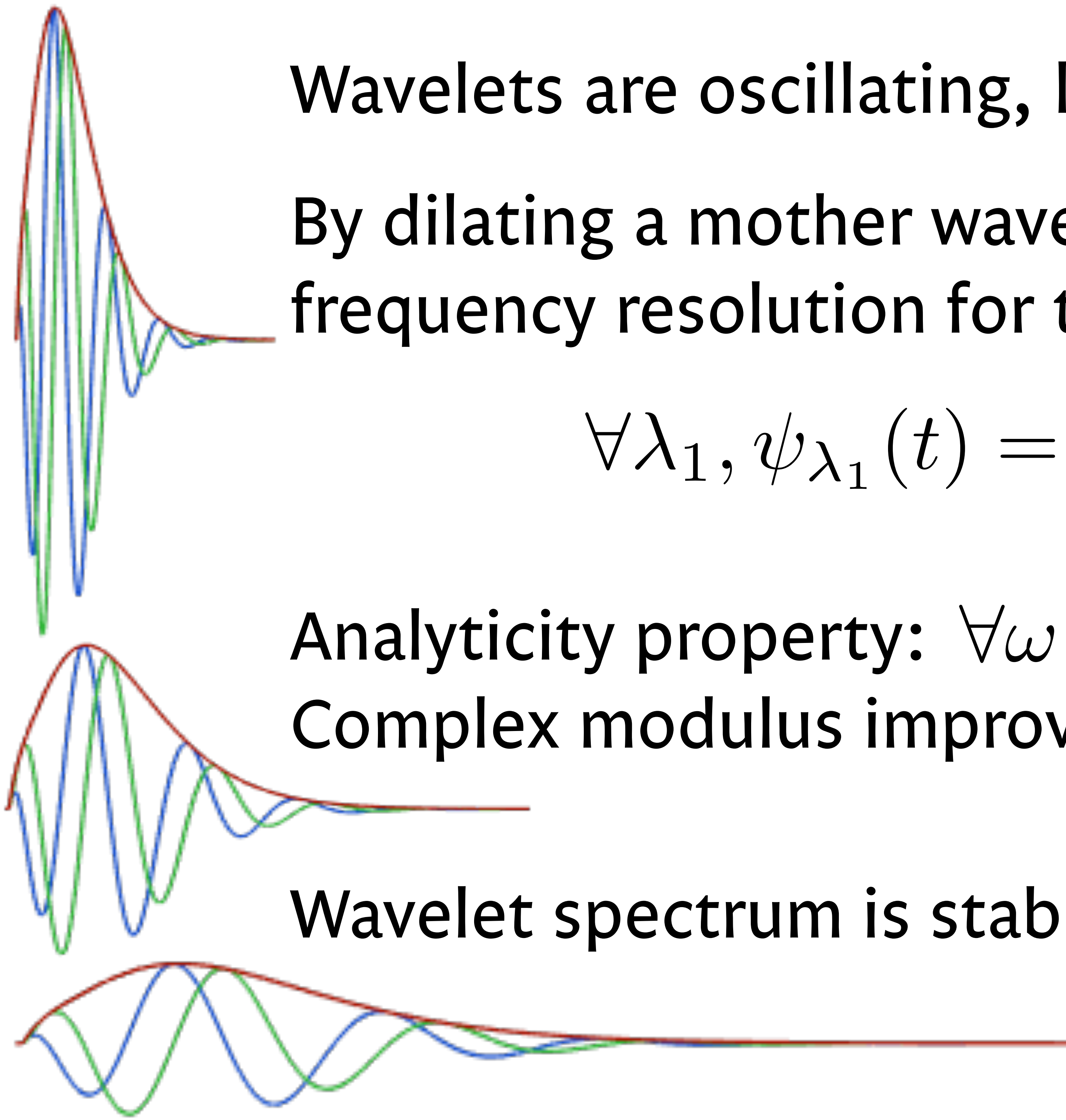
We need a time-frequency representation that is

1. localized enough in time
   (a) to have $\dot{\theta}(t)$ approximately constant,
   (b) to have $\dot{\nu}(t)$ approximately constant,

2. localized enough in frequency
   (a) to distinguish the first peaks of $\hat{e}(\omega)$,
   (b) to have $\hat{h}(\omega)$ approximately constant.

We will use a constant-Q filter bank of wavelets.

Analytic wavelets



Wavelets are oscillating, localized filters.

By dilating a mother wavelet $\psi(t)$, we trade frequency resolution for time resolution.

$$\forall \lambda_1, \psi_{\lambda_1}(t) = \lambda_1 \psi(\lambda_1 t)$$

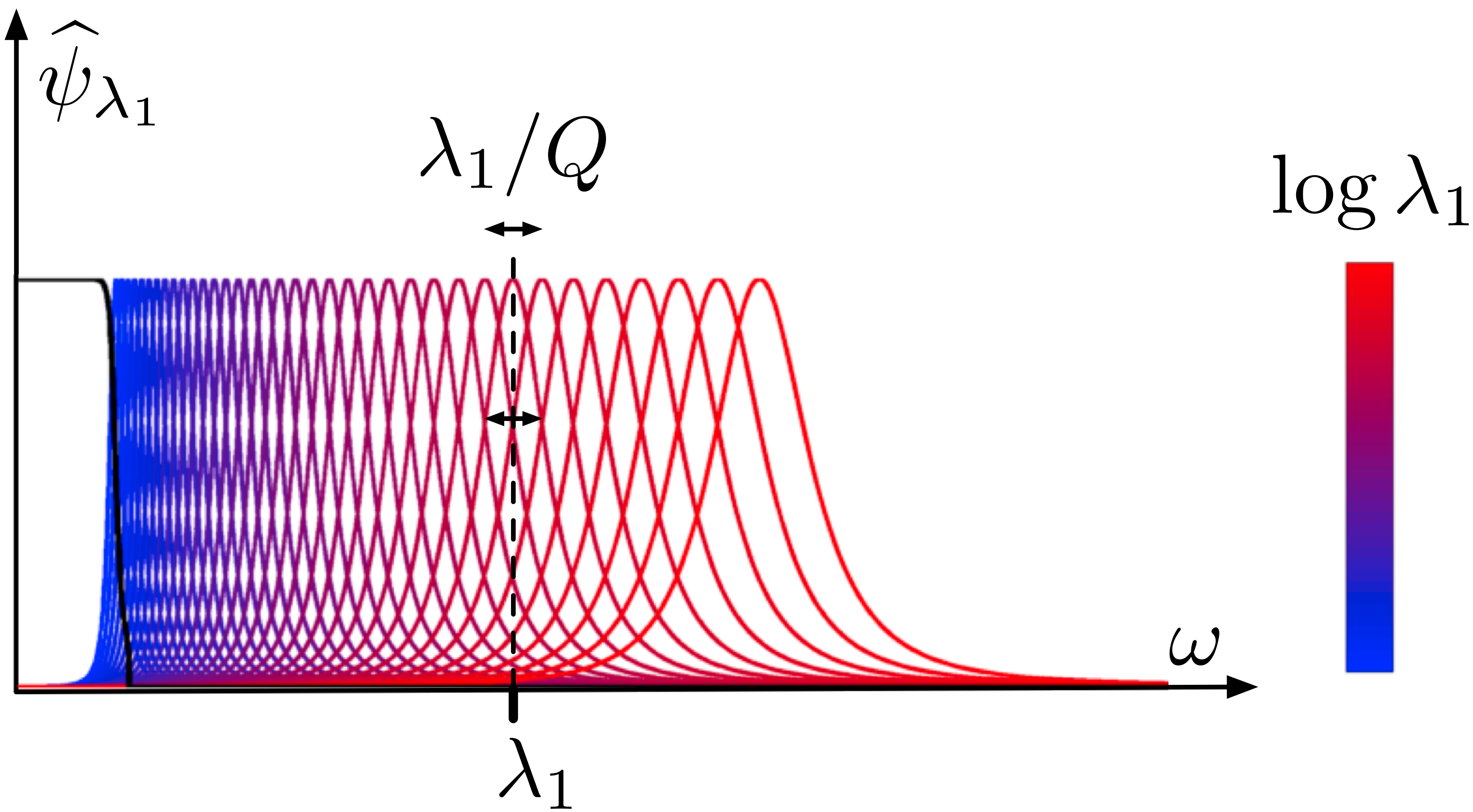
Analyticity property: $\forall \omega \leq 0, \widehat{\psi}(\omega) = 0.$
Complex modulus improves regularity.

Wavelet spectrum is stable to deformations.

Wavelet filter bank

In base 2: $\log \lambda_1 = j_1 + \frac{\chi_1}{Q}$ **number of filters per octave**

where the integer part j_1 is the octave index and $\chi_1 \in \{0 \ldots (Q-1)\}$ is the chroma.



Wavelet ridge theorem

by Delprat, Escudié, Guillemain, Kronland-Martinet, Torrésani, and Tchamitchian (1992).

Let $f(t) = a(t)\cos\theta(t)$ and $\psi_{\lambda_1}(t) = \lambda_1 g(\lambda_1 t)\exp(\mathrm{i}\lambda_1 t)$.

$$\boxed{\begin{aligned}&[f * \psi_{\lambda_1}](t)\\ &\quad = a(t)\exp(\mathrm{i}(\theta(t) - \lambda_1 t)) \times \left(\hat{g}\left(1 - \frac{\dot{\theta}(t)}{\lambda_1}\right) + \varepsilon(t, \lambda_1)\right)\end{aligned}}$$

The corrective term $\varepsilon(t, \lambda_1)$ is small if

- amplitude modulation is slow: $\|\dot{a}/a\|_\infty^2, \|\ddot{a}/a\|_\infty \ll \lambda_1^2$,
- frequency modulation is slow: $\|\ddot{\theta}\|_\infty \ll \lambda_1^2$, and
- (t, λ_1) is near a ridge: $\dot{\theta}(t) \approx \lambda_1$.

Factorization in the scalogram

For Q between 12 and 24, we have

$$|x_{\theta,\nu} * \psi_{\lambda_1}| = \widehat{\psi}_{\lambda_1}(k\dot{\theta}(t)) \times \hat{h}\left(\frac{\lambda_1}{\dot{\nu}(t)}\right)$$

where k is such that $\lambda_1 \approx k\dot{\theta}(t)$.

In log-frequency and after logarithmic compression:

$$\mathbf{U_1} x_{\theta,\nu}(t, \log \lambda_1) \overset{\text{def.}}{=} \log |x_{\theta,\nu} * \psi_{\lambda_1}|$$

$$= \underbrace{\mathbf{U_1} e(\log \lambda_1 - \log \dot{\theta}(t))}_{\text{translated source}} + \underbrace{\mathbf{U_1} h(\log \lambda_1 - \log \dot{\nu}(t))}_{\text{translated filter}}$$

2. Leveraging harmonicity

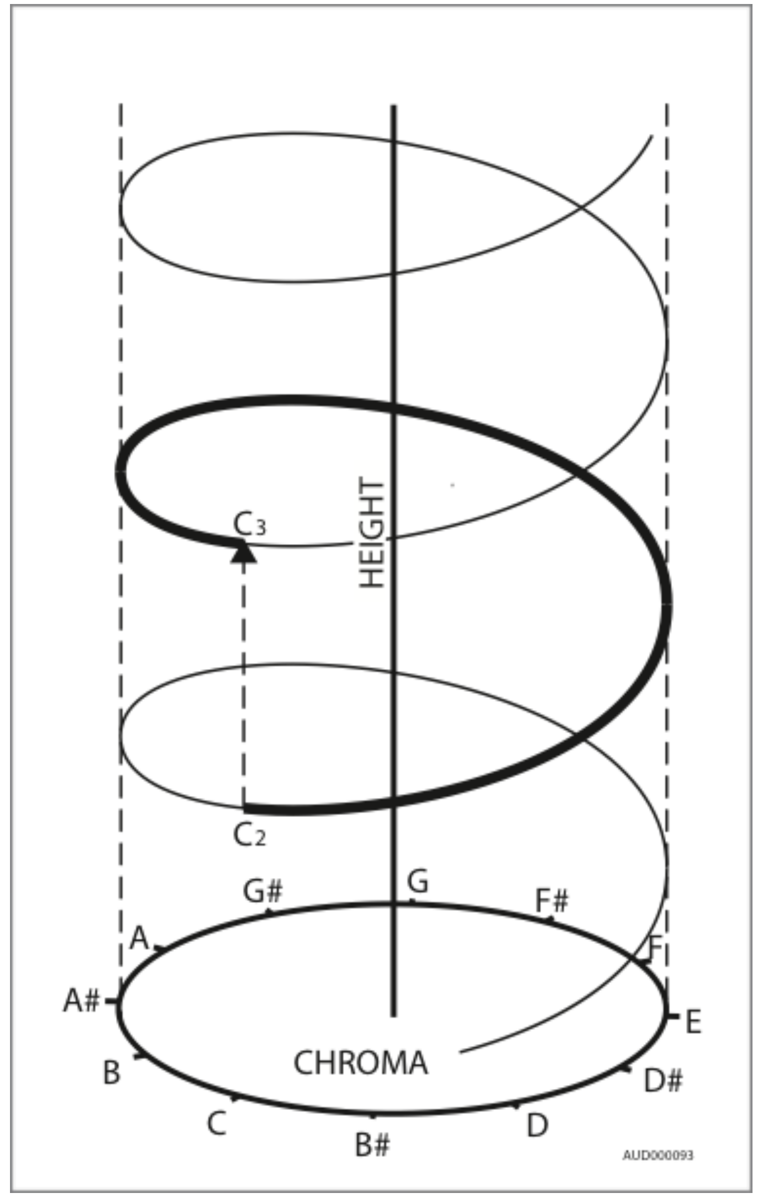


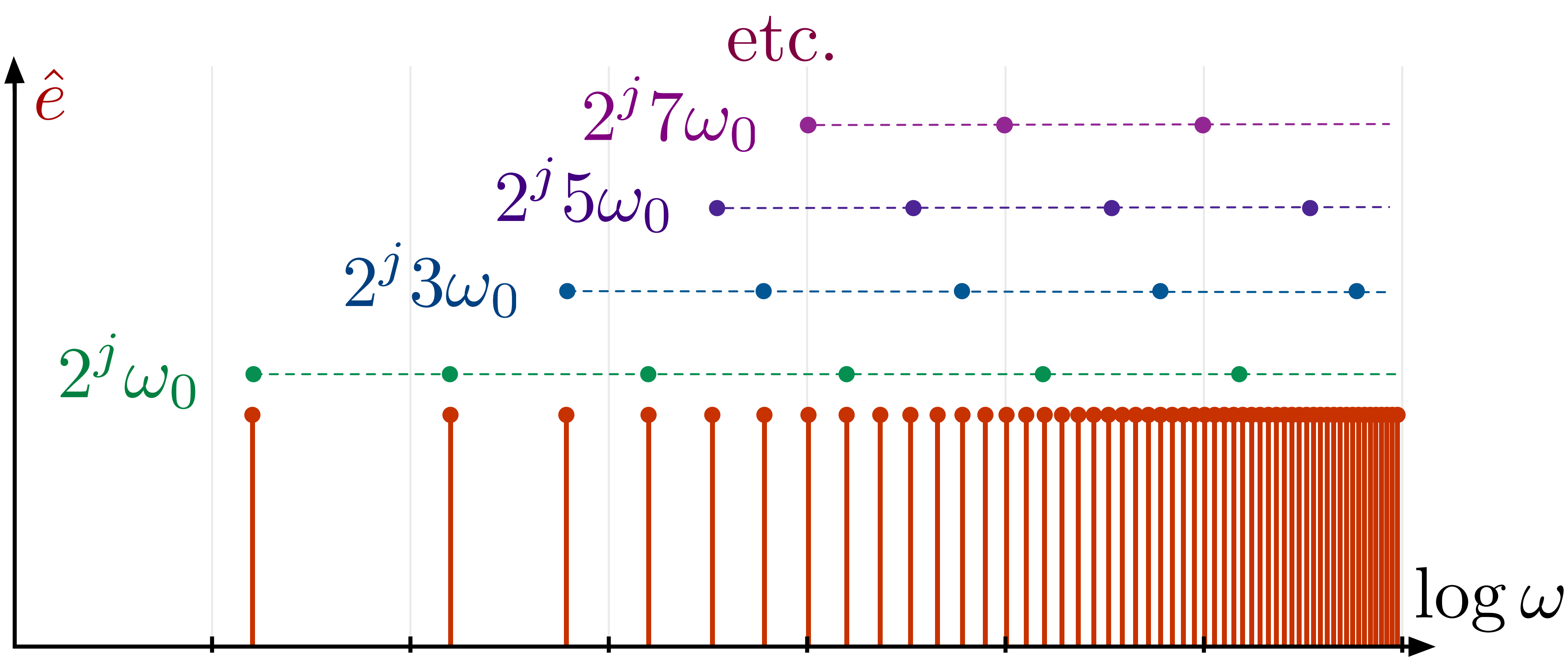
Fig. from Shepard (1964).

HARMONICITY PROPERTY

The harmonic comb is self-similar:

$$\hat{e}(\omega) = \hat{e}(2^j \omega) \text{ for all } \omega > 1 \text{ and } j \in \mathbb{N}.$$

Regularity across octaves for a given chroma:

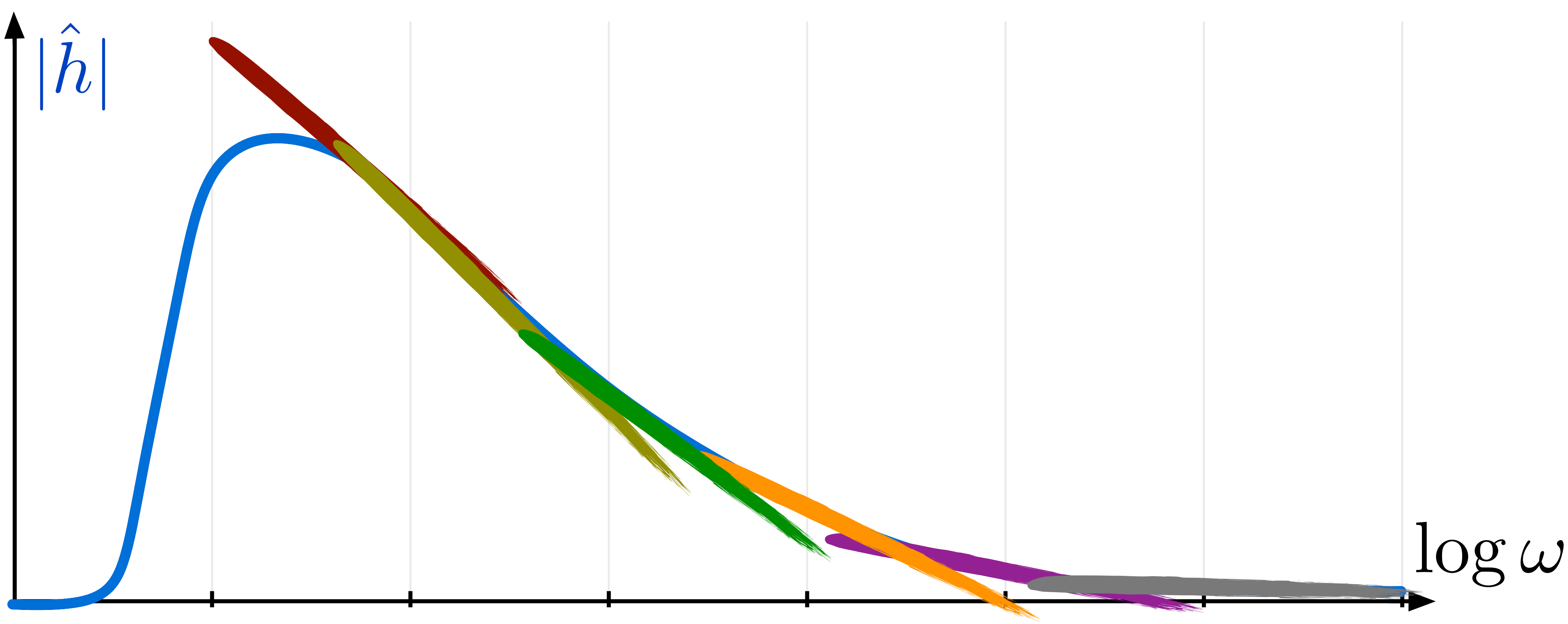


Spectral smoothness property

The spectral envelope is regular across semitones:

$$\frac{\partial \log |\hat{h}|}{\partial \omega} \ll \frac{\partial \log |\hat{e}|}{\partial \omega}.$$

Regularity along chromas within an octave:



PARTIAL DERIVATIVE ALONG TIME

By linearity and chain rule formula:

$$\frac{\partial \mathbf{U_1} x_{\theta,\nu}}{\partial t}(t, \log \lambda_1)$$

$$= \frac{\ddot{\theta}(t)}{\dot{\theta}(t)} \frac{\mathrm{d}\mathbf{U_1} e}{\mathrm{d}(\log \lambda_1)}(\log \lambda_1 - \log \dot{\theta}(t))$$

$$+ \frac{\ddot{\nu}(t)}{\dot{\nu}(t)} \frac{\mathrm{d}\mathbf{U_1} h}{\mathrm{d}(\log \lambda_1)}(\log \lambda_1 - \log \dot{\nu}(t)).$$

Partial derivative along chromas

By linearity:

$$\frac{\partial \mathbf{U_1} x_{\theta,\nu}}{\partial(\log \lambda_1)}(t, \log \lambda_1) = \frac{\mathrm{d}\mathbf{U_1} e}{\mathrm{d}(\log \lambda_1)}(\log \lambda_1 - \log \dot{\theta}(t)) + \underbrace{\frac{\mathrm{d}\mathbf{U_1} h}{\mathrm{d}(\log \lambda_1)}(\log \lambda_1 - \log \dot{\nu}(t))}_{\text{neglected because of spectral smoothness}}.$$

Partial derivative across octaves

By linearity:

$$\frac{\Delta \mathbf{U_1} x_{\theta,\nu}}{\Delta j_1}(t, \log \lambda_1) = \underbrace{\frac{\Delta \mathbf{U_1} e}{\Delta j_1}(\log \lambda_1 - \log \dot{\theta}(t))}_{\text{neglected because of } \textbf{harmonicity}}$$

$$+ \frac{\Delta \mathbf{U_1} h}{\Delta j_1}(\log \lambda_1 - \log \dot{\nu}(t)).$$

Optical flow equation

$\mathbf{U_1}x$ behaves like an object in rigid motion,
hence an optical flow equation in t, $\log \lambda_1$, and j_1.

$$\frac{\partial \mathbf{U_1} x_{\theta,\nu}}{\partial t}(t, \log \lambda_1) = \frac{\ddot{\theta}(t)}{\dot{\theta}(t)} \frac{\partial \mathbf{U_1} x_{\theta,\nu}}{\partial(\log \lambda_1)}(t, \log \lambda_1)$$

$$+ \frac{\ddot{\nu}(t)}{\dot{\nu}(t)} \frac{\Delta \mathbf{U_1} x_{\theta,\nu}}{\Delta j_1}(t, \log \lambda_1).$$

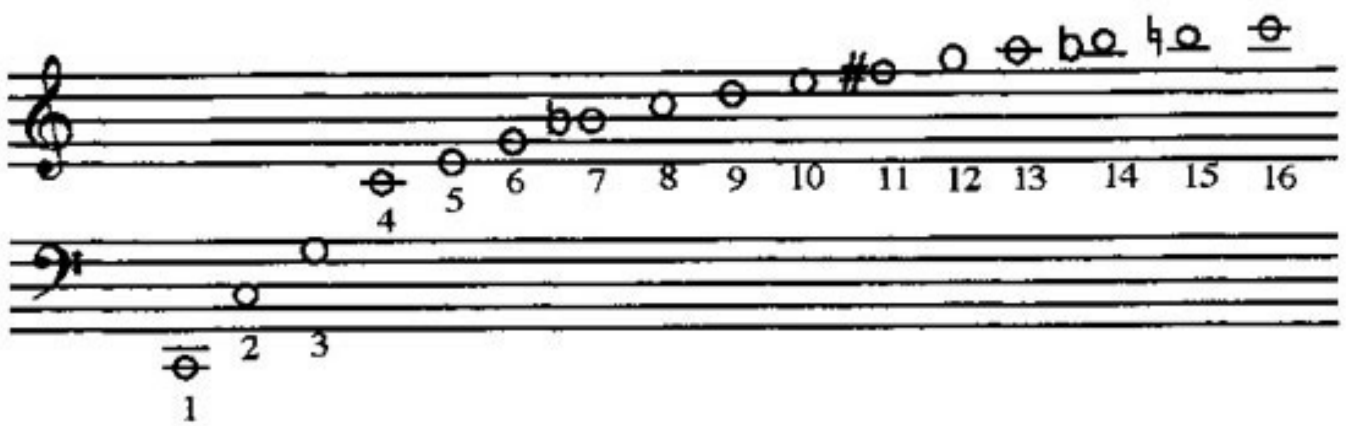
Motion in $(\log \lambda_1, j_1)$ is best expressed on a **spiral**:
the chroma is angular, the octave is radial.

SHEPARD-RISSET HELIX

Two continuous paths from C2 to C3

- « staircase » : *chirp*

- « lift » : attenuation of odd-numbered partials



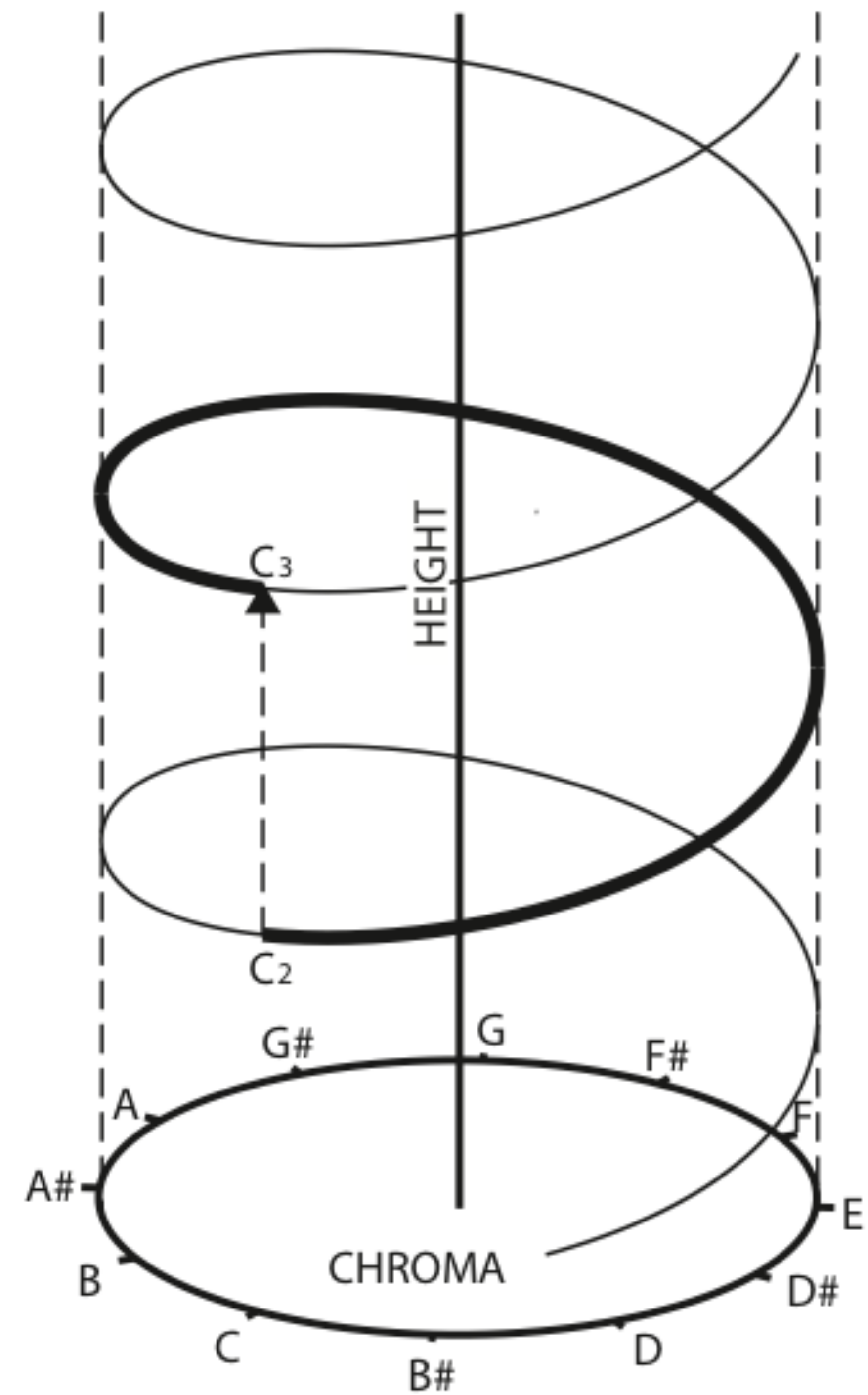


Fig. from Shepard (1964).

We replace the 3D helix by a 2D spiral.

ROTATING MOTION IN THE SPIRAL

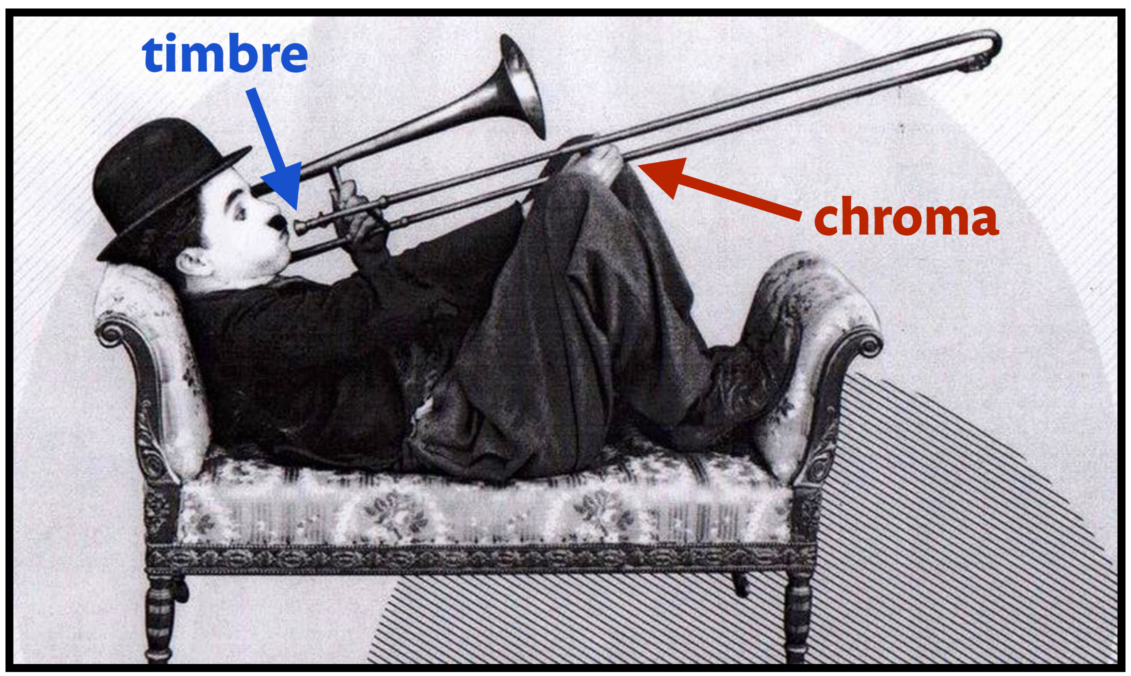
A chirp has a rotating motion in the spiral.

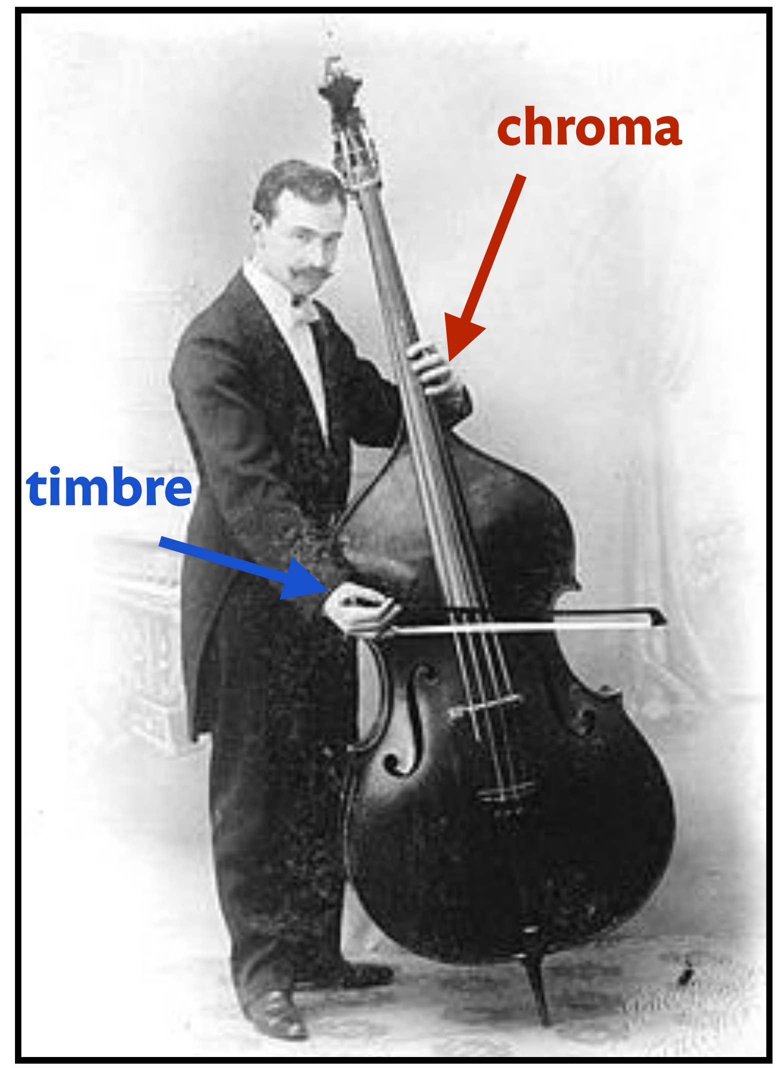
RADIAL MOTION IN THE SPIRAL

A formantic change has a radial motion in the spiral.

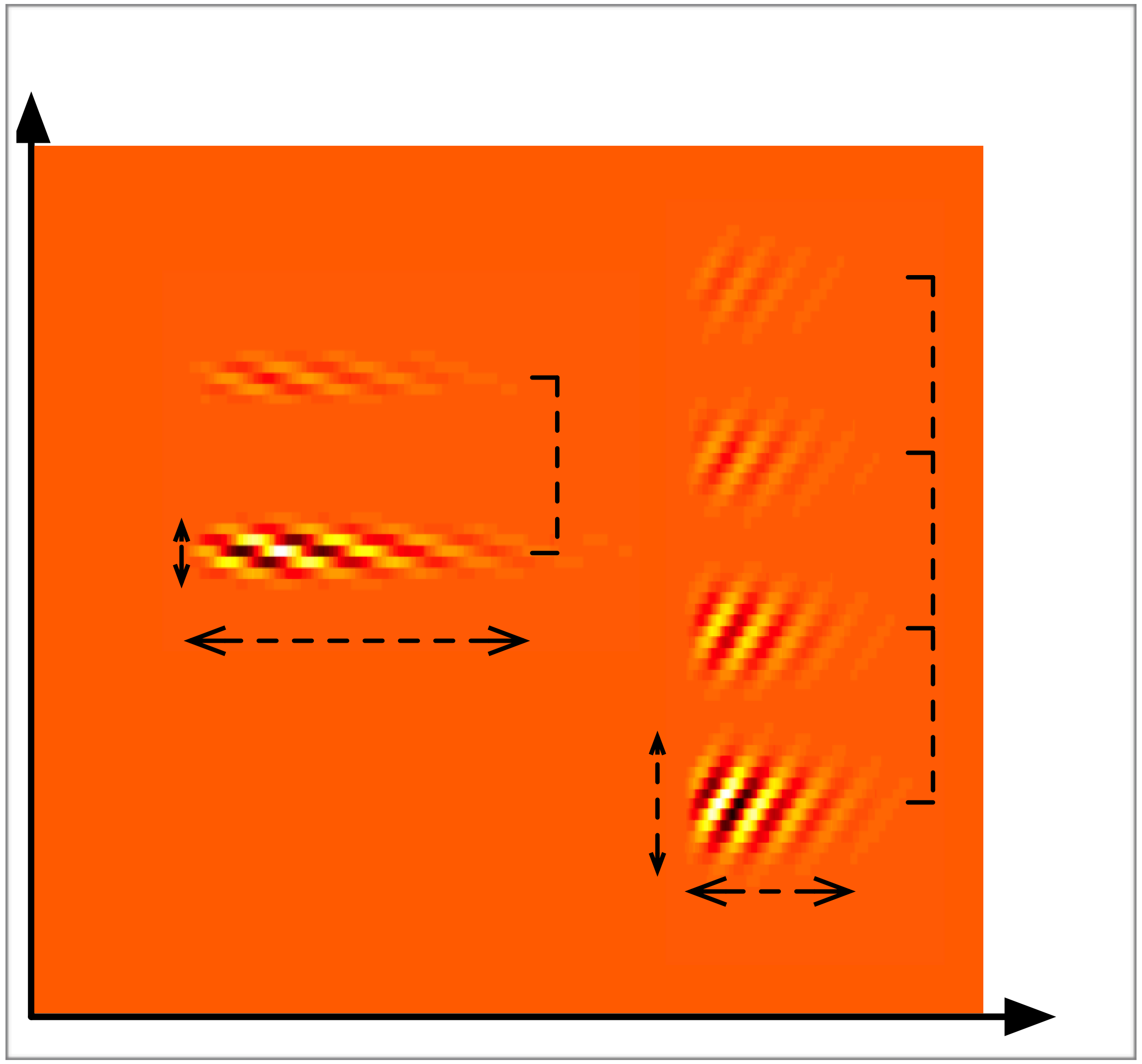
Two degrees of freedom

Musical transients are not regular in time-frequency...
... but in **time-chroma-octave**.





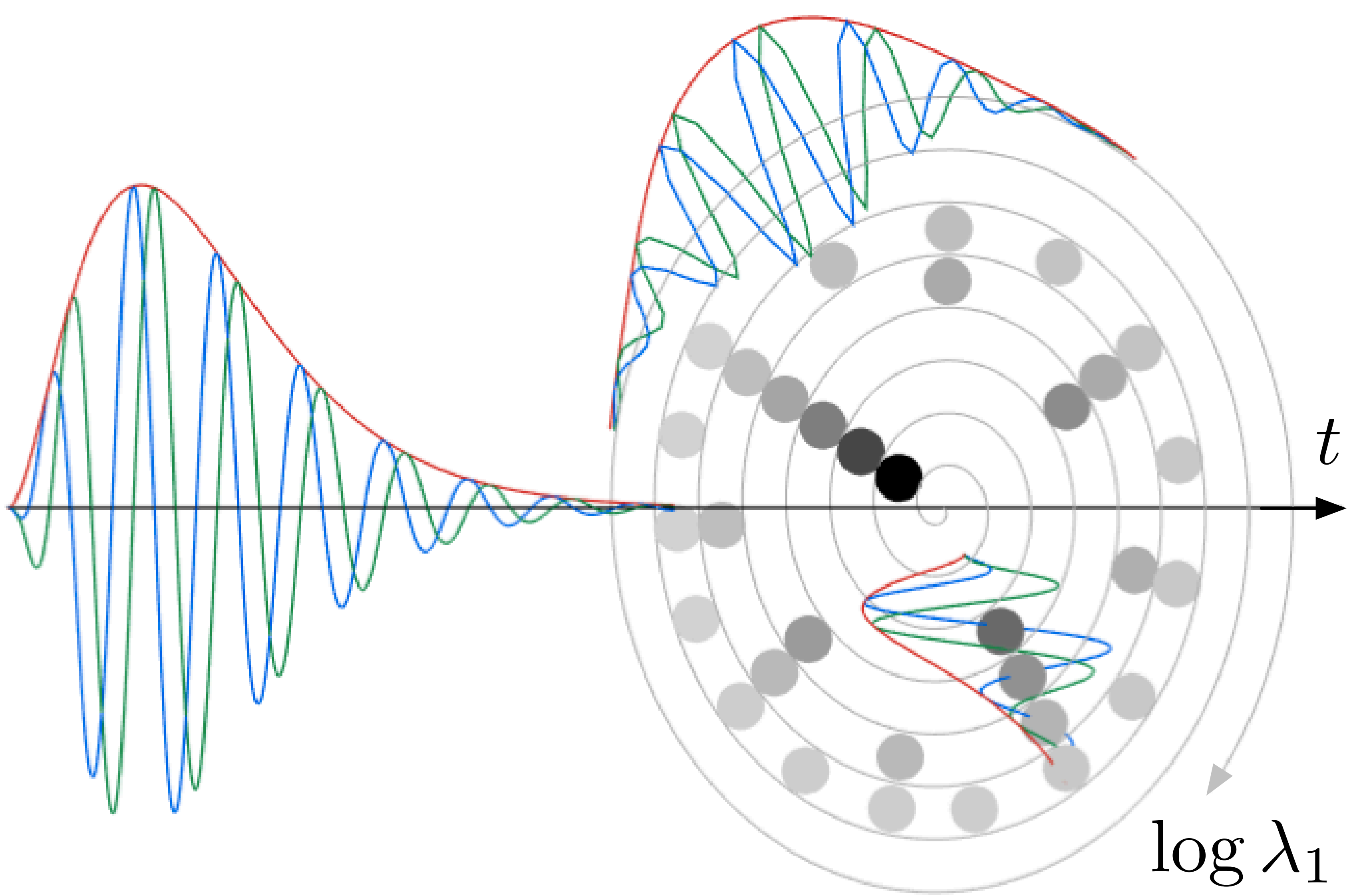
3. Spiral scattering



Spiral wavelets

Accepted at GRETSI 2015 with Stéphane Mallat:
Transformée de scattering en spirale temps-chroma-octave.

$$\Psi_{(\alpha,\beta,\gamma)}(t) = \psi_\alpha(t) \times \psi_\beta(\log \lambda_1) \times \psi_\gamma(\lfloor \log \lambda_1 \rfloor)$$



Frequencies vs. Quefrencies

$$\Psi_{(\alpha,\beta,\gamma)}(t) = \psi_\alpha(t) \times \psi_\beta(\log \lambda_1) \times \psi_\gamma(\lfloor \log \lambda_1 \rfloor).$$

- α is a modulation frequency along time, in Hertz.
  α^{-1} is typically between 1 ms and 100 ms.

- β is a « quefrency » along chromas, in cycles per octaves.
  $|\beta^{-1}|$ is typically between 1 semitone and 1 octave.

- γ is a quefrency across octaves, in cycles per octaves.
  $|\gamma^{-1}|$ is typically between 1 and 4 octaves.

We define the multiindex frequency variable $\lambda_2 = (\alpha, \beta, \gamma)$.
By convention, $\log \lambda_2 = (\log \alpha, \log \beta, \mathrm{sign}\ \beta, \log \gamma, \mathrm{sign}\ \gamma)$.

SCATTERING CASCADE

Wavelet filter banks *scatter* the energy from $\mathbf{U_m}$ to $\mathbf{U_{m+1}}$.

Complex modulus improves regularity and phase invariance.

$$\mathbf{U_1}x(t, \log\lambda_1) = |x \overset{t}{*} \psi_{\lambda_1}|(t)$$

$$\mathbf{U_2}x(t, \log\lambda_1, \log\lambda_2) = |\mathbf{U_1}x \circledast \Psi_{\lambda_2}|(t, \log\lambda_1)$$

multivariable convolution

A lowpass filter $\phi(t)$ enforces the amount of translation invariance that is required by the classification task.

$$\mathbf{S_1}x(t, \log\lambda_1) = \mathbf{U_1}x \overset{t}{*} \phi$$

$$\mathbf{S_2}x(t, \log\lambda_1) = \mathbf{U_2}x \overset{t}{*} \phi$$

Source-filter properties

- **Vanishing moment** property:
  Convolving a wavelet with a linear function yields almost zero.

- Harmonicity and spectral smoothness rewrite as

$$\left| \mathbf{U_1} e_\theta \overset{j_1}{*} \psi_\gamma \right| = 0 \quad \text{and} \quad \left| \mathbf{U_1} h_\nu \overset{\chi_1}{*} \psi_\beta \right| \approx 0.$$

- The spiral scattering transform boils down to

$$\left| \mathbf{U_1} x_{\theta,\nu} \overset{t,\chi_1,j_1}{\circledast} \Psi_{\lambda_2} \right|$$

$$= \left[\left(\mathbf{U_1} e_\theta \overset{\chi_1}{*} \psi_\beta \right) \times \left(\mathbf{U_1} h_\nu \overset{j_1}{*} \psi_\gamma \right) \right] \overset{t}{*} \psi_\alpha$$

Spiral wavelet ridges

- Applying the wavelet ridge theorem three times yields:

$$\left| \mathbf{U_1} x_{\theta,\nu} \overset{t,\chi_1,j_1}{\circledast} \Psi_{\lambda_2} \right|$$
$$= \left| \mathbf{U_1} e_\theta \overset{\chi_1}{*} \psi_\beta \right| \left| \mathbf{U_1} h_\nu \overset{j_1}{*} \psi_\gamma \right| \left| \widehat{\psi}_\alpha \left(-\frac{\ddot{\theta}(t)}{\dot{\theta}(t)} \beta - \frac{\ddot{\nu}(t)}{\dot{\nu}(t)} \gamma \right) \right|$$

- Ridges are on a plane whose Cartesian equation is

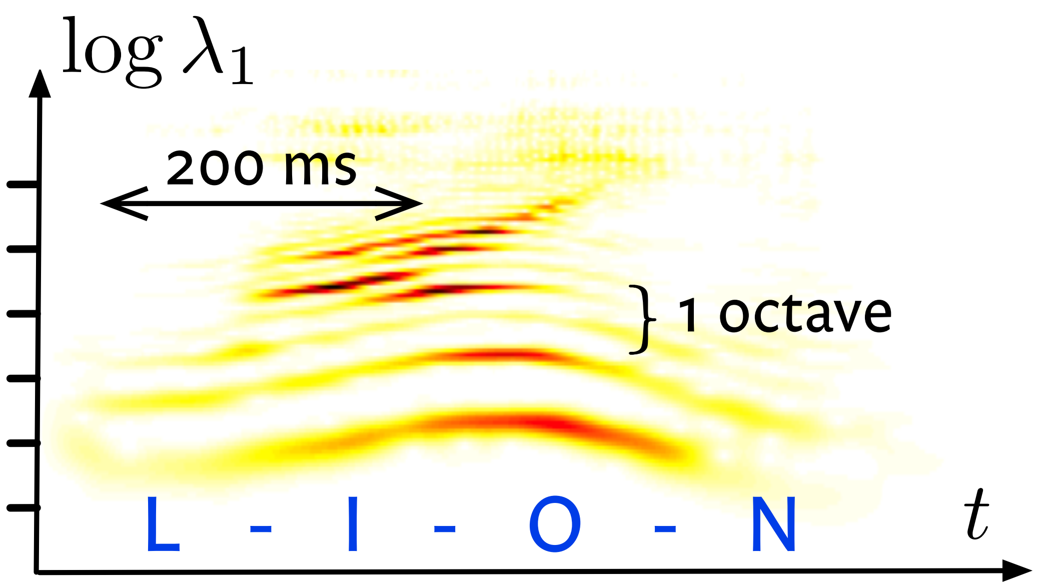
$$\alpha + \frac{\ddot{\theta}(t)}{\dot{\theta}(t)} \beta + \frac{\ddot{\nu}(t)}{\dot{\nu}(t)} \gamma = 0\,.$$

- The same holds for averaged coefficients $\mathbf{S_2} x_{\theta,\nu}$ over T if

$$\left| \frac{\dddot{\theta}(t)}{\ddot{\theta}(t)} - \frac{\ddot{\theta}(t)}{\dot{\theta}(t)} \right| \ll T^{-1} \quad \text{and} \quad \left| \frac{\dddot{\nu}(t)}{\ddot{\nu}(t)} - \frac{\ddot{\nu}(t)}{\dot{\nu}(t)} \right| \ll T^{-1}.$$

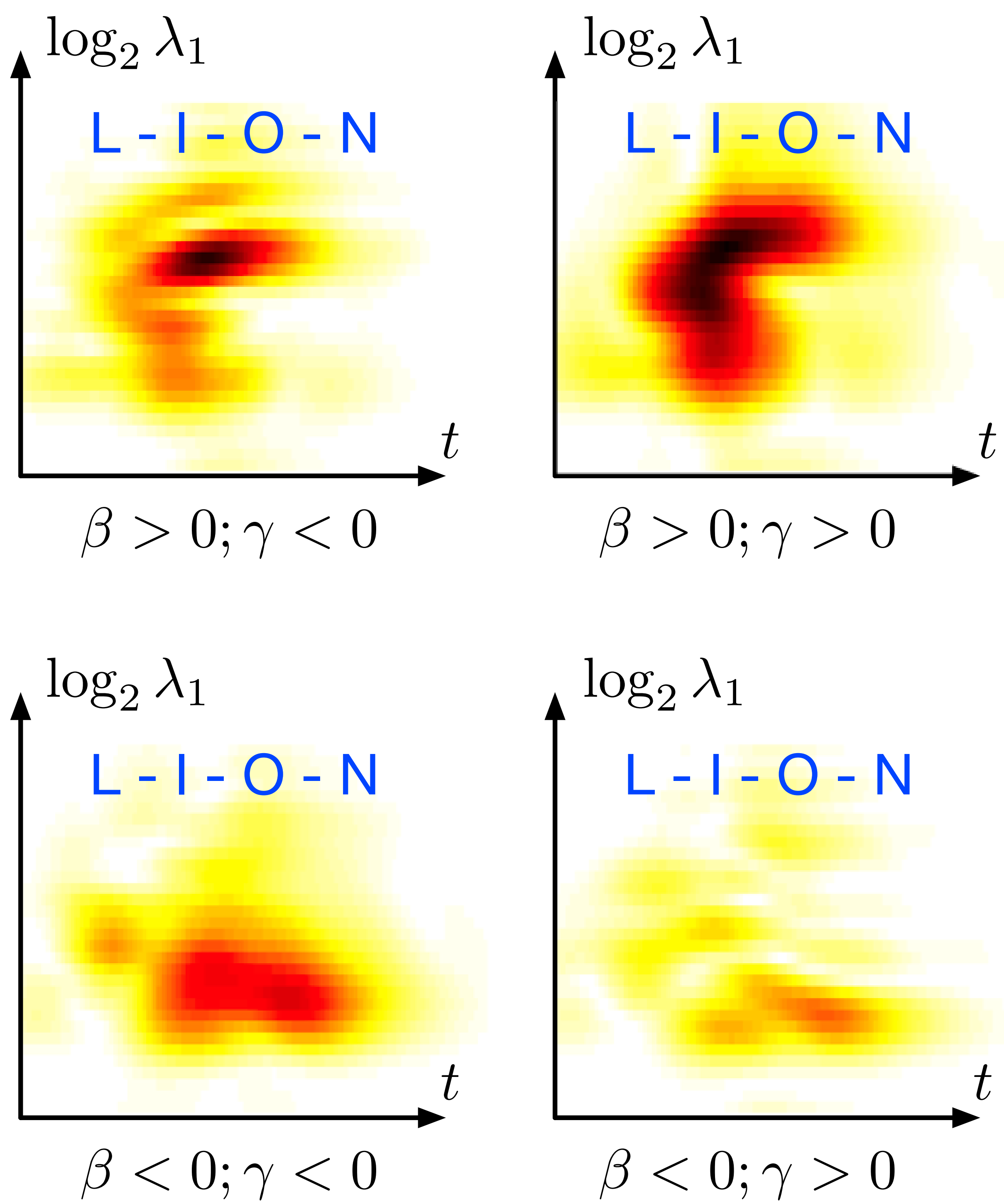
Spatial localization of ridges

Below:
$\mathbf{U_1} x$ of the word « lion ».



Opposite: $\mathbf{U_2} x$ slices for

- $\alpha^{-1} = 120$ ms
- $|\beta^{-1}| = \pm 1$ octave
- $|\gamma^{-1}| = \pm 1$ octave



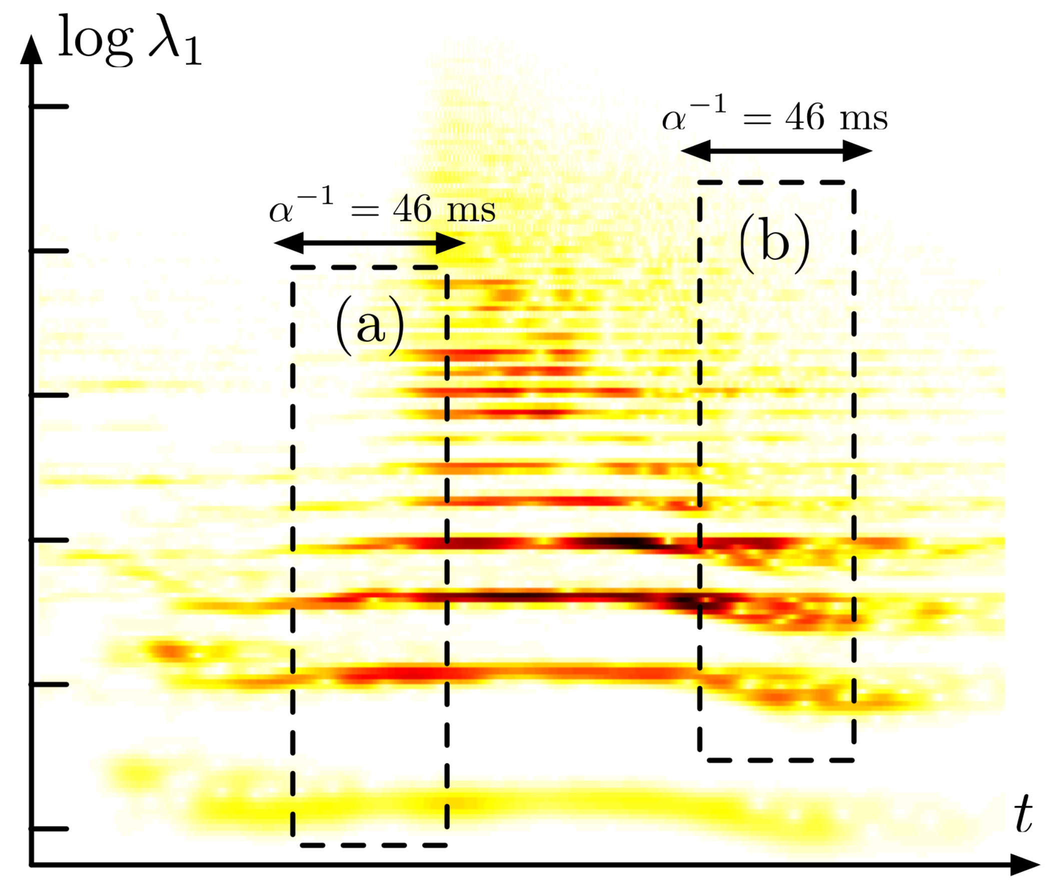
(α, β, γ) LOCALIZATION OF RIDGES

Top:

$\mathbf{U_1}x$ of a trombone note.

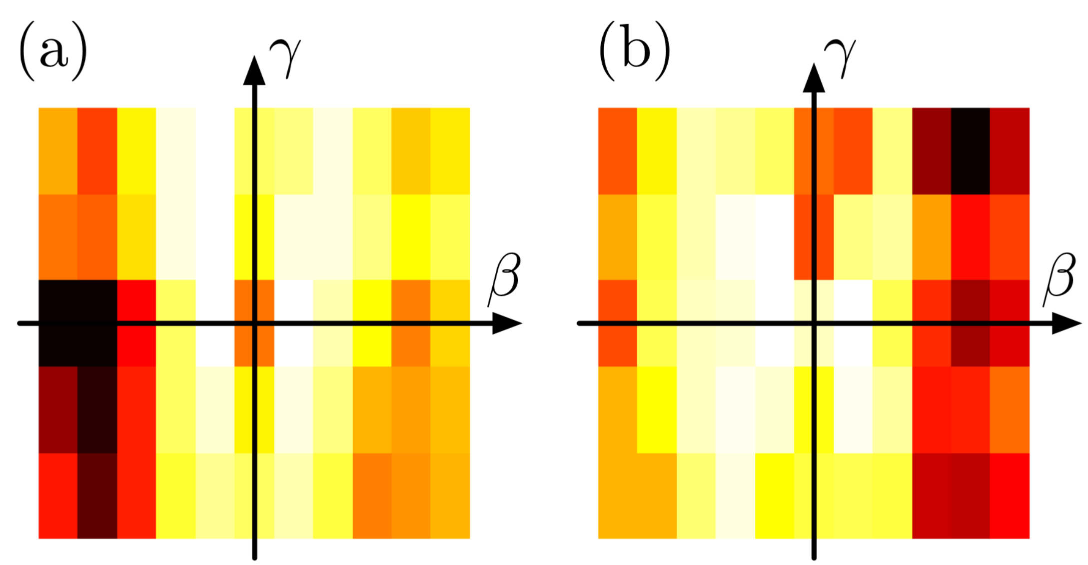
(a) **attack** part with upwards glissando

(b) **release** part with downwards glissando.



Bottom: $\mathbf{U_2}x$ slices for

- $\alpha^{-1} = 46$ ms
- fixed t and $\log \lambda_1$.



PHONEME CLASSIFICATION

Submitted to MLSP 2015 with Joakim Andén and Stéphane Mallat:
Joint Time-frequency Scattering for Audio Classification.

Phoneme error rate on the TIMIT dataset [Fisher et al. 1986].

MFCC and SVM	18,3 %
MFCC and GMM commitee [Chang & Glass, 2007]	16.7 %
α scattering and SVM	17,3 %
(α, β) scattering and SVM	15,8 %
(α, β, γ) scattering	coming next

INVARIANT RECONSTRUCTION

Problem:
« find a translation-invariant representation that allows the most plausible signal reconstruction ».

Given a target $\mathbf{S}^\infty = (\mathbf{S}_\mathbf{1}^\infty, \mathbf{S}_\mathbf{2}^\infty)$, the gradient descent

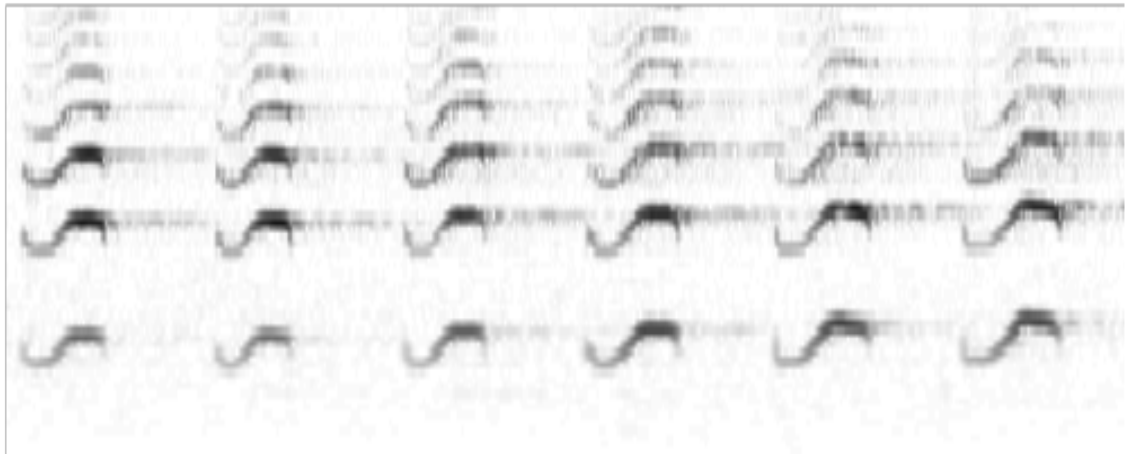
$$\Delta \mathbf{U}_\mathbf{1} x = \Delta \mathbf{S}_\mathbf{1} x \overset{t}{*} \bar{\phi}$$

$$+\Re \left[\sum_{\lambda_2} \left(\frac{\mathbf{U}_\mathbf{1} x \circledast \Psi_{\lambda_2}}{|\mathbf{U}_\mathbf{1} x \circledast \Psi_{\lambda_2}|} \times \Delta \mathbf{U}_\mathbf{2} x \right) \circledast \bar{\Psi}_{\lambda_2} \right].$$

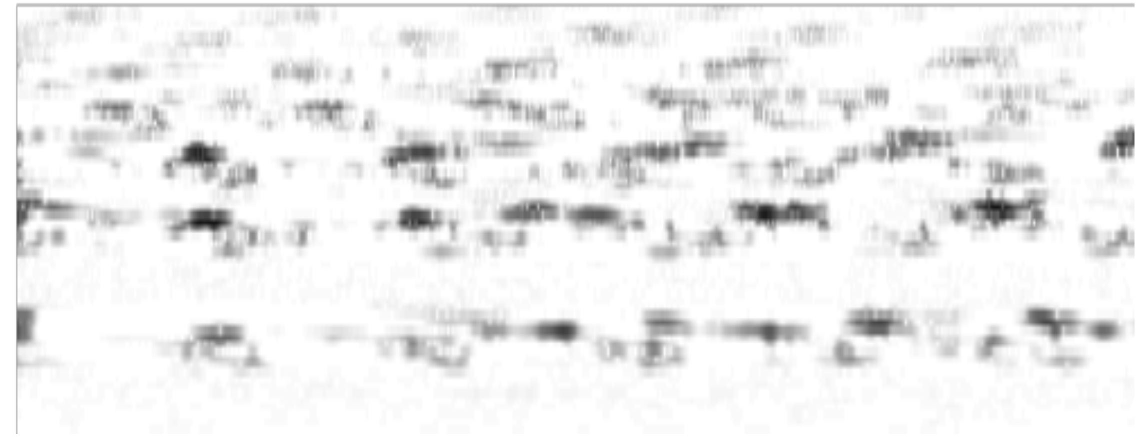
converges to a local minimum of the loss function

$$E(x) = \|\mathbf{S}_\mathbf{1} x - \mathbf{S}_\mathbf{1}^\infty\|_2 + \|\mathbf{S}_\mathbf{2} x - \mathbf{S}_\mathbf{2}^\infty\|_2 .$$

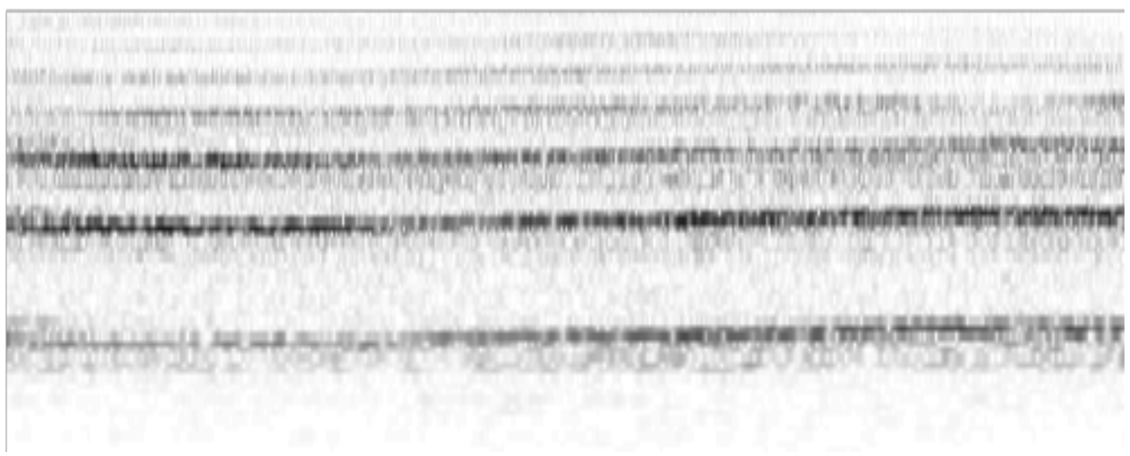
original



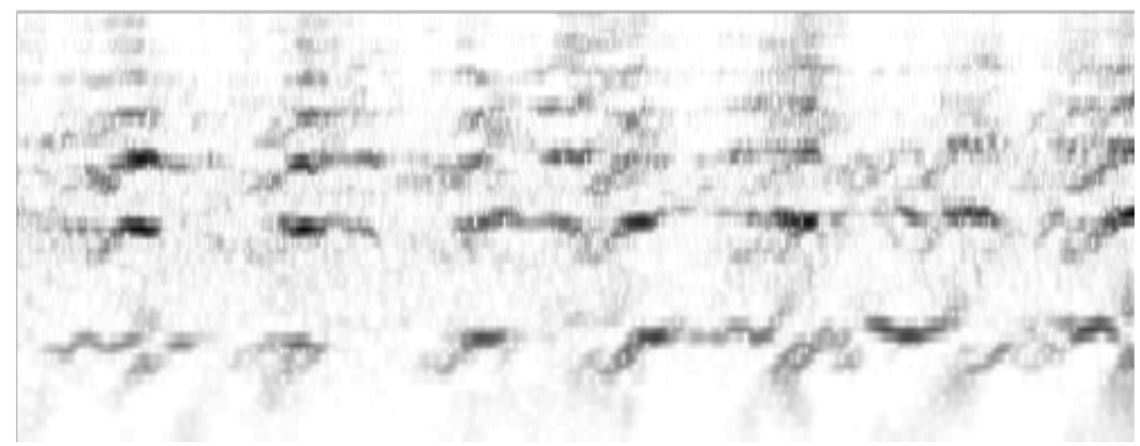
time scattering



first-order only

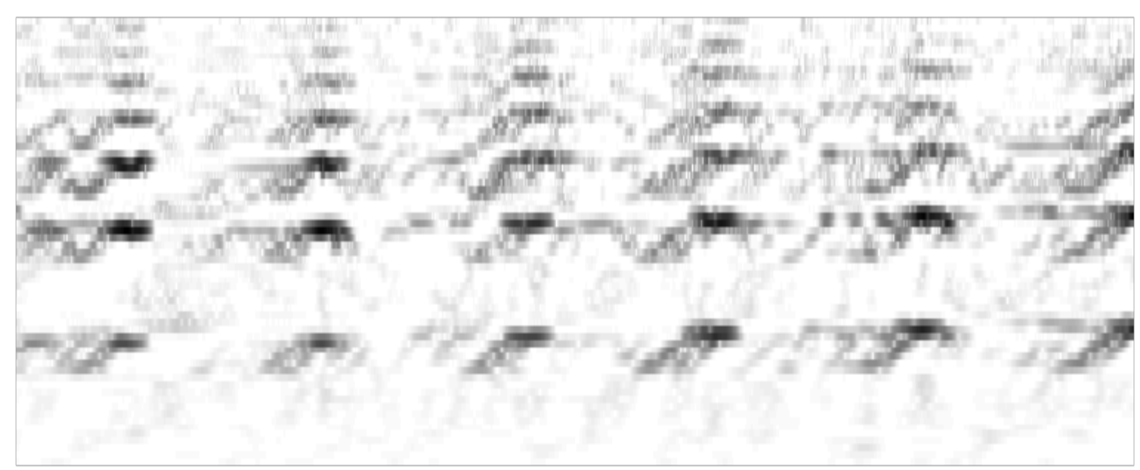


time-frequency scattering



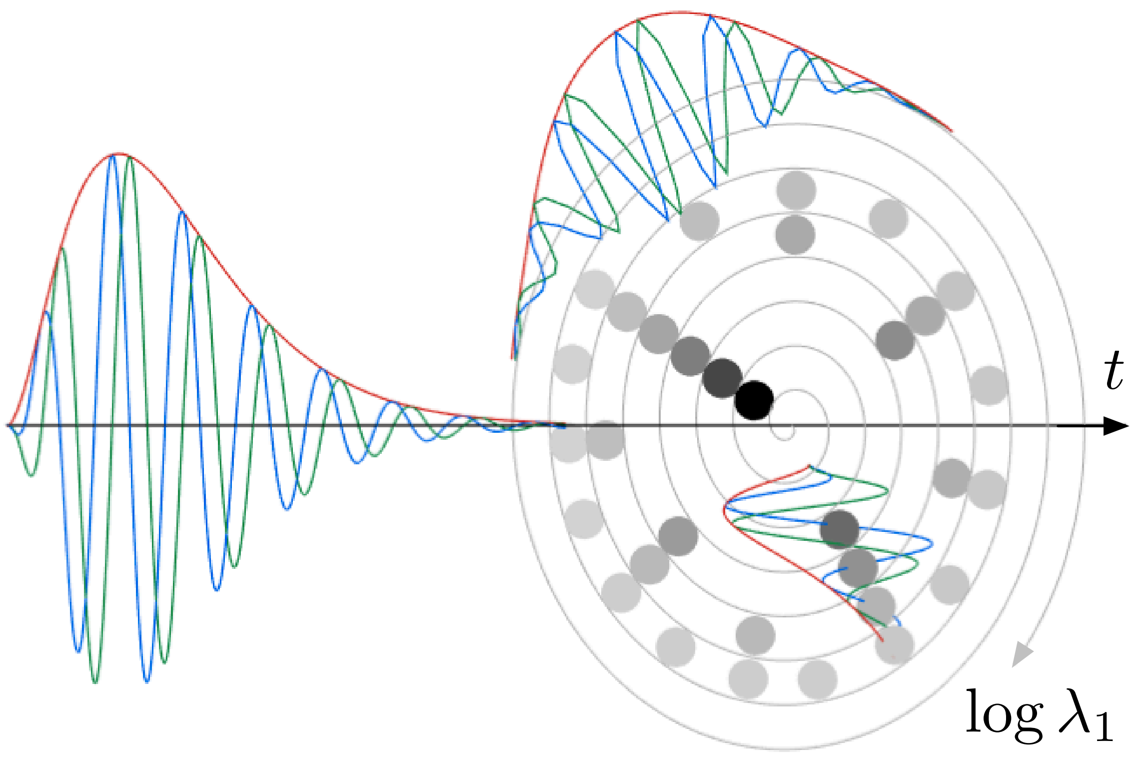
McDermott & Simoncelli

spiral scattering



Conclusions

- Natural sounds are nonstationary, but physically regular.
- In the pitch spiral, source-filter transients become translations.
- Spiral scattering yields source-filter velocities **without detection**.
- Encouraging results in classification and invariant reconstruction.



Experiments can be reproduced at:
www.github.com/lostanlen/