WAVELET SCATTERING ON THE PITCH SPIRAL

*Vincent Lostanlen, Stéphane Mallat**

Dept. of Computer Science,
École normale supérieure
Paris, France
vincent.lostanlen@ens.fr

ABSTRACT

We present a new representation of sounds that liearizes the dynamics of pitch chroma and pitch height, while remaining stable to deformations in the time-frequency plane. It is an instance of the scattering transform, a generic operator which cascades wavelet convolutions and modulus nonlinearities. It is derived from the Shepard pitch spiral, in that convolutions are performed in time, log-frequency (correlated to pitch chroma) and octave index (correlated to pitch height).

1. INTRODUCTION

Understanding the natural regularities in audio signals is essential to the design of a useful representation for classification, blind source separation, automated transcription, as well as other processing tasks. It is well known that a wavelet decomposition followed by complex modulus brings short-term regularity along the time axis, as it locally replaces oscillatory components by their envelopes. After this operation, longer-term regularities appear. In this article, we show that the same "scattering" scheme of wavelet convolutions of complex modulus, can be applied to encompass them.

First, local continuity in frequency modulation cause a slow rigid motion along the log-frequency axis. Second, harmonic sounds exhibit a comb-like spectrum, which conveys the global evolution of the spectral envelope. This comb looks highly irregular: unlike frequency modulation, it cannot be captured efficiently with local convolutions in time and log-frequency.

To recover regularity across partials, we capitalize on the fact that power-of-two harmonics are distant from exactly one octave. By rolling up the log-frequency axis in a spiral, such that octave intervals correspond to full turns, these partials get aligned on a radius. Consequently, introducing the integer-valued octave variable reveals harmonic regularity that was not explicit in the plane of time and log-frequency. Once specified the variables of time, log-frequency, and octave index, our transform merely consists in cascading three wavelet decompositions along them and applying complex modulus.

Section 2 gives a formal definition of the spiral scattering transform. Section 3 introduces a nonstationary formulation of the source-filter model relying on time warps, and shows that its variabilities in pitch and spectral envelope are jointly linearized by the spiral scattering transform. Section 4 provides a visual interpretation of the spiral wavelet coefficients in a music signal with extended instrumental techniques.

* This work is supported by the ERC InvariantClass 320959.

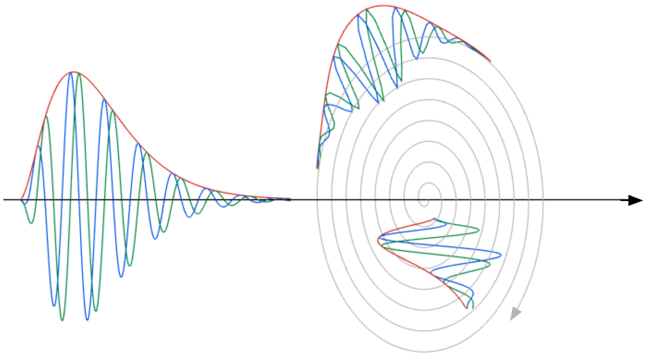

Figure 1

2. FROM TIME SCATTERING TO SPIRAL SCATTERING

2.1. Time scattering

We denote the Fourier transform of a signal $x(t)$ by $\widehat{x}(\omega) = \int x(u) \exp(-\mathrm{i}\omega u)$
An analytic mother wavelet is a complex filter $\psi(t)$ whose Fourier transform $\widehat{\psi}(\omega)$ is concentrated over the frequency interval $[1 - 2^{1/2Q}, 1 + 2^{1/2Q}]$. Dilations of this wavelet filter defines a family of filters centered at frequencies $\lambda_1 = 2^{j_1 + \frac{\chi_1}{Q}}$ where the indices $j_1 \in \mathbb{Z}$ and $\chi_1 \in \{1 \ldots Q\}$ respectively denote octave and chroma:

$$\widehat{\psi}_{\lambda_1}(\omega) = \widehat{\psi}(\lambda^{-1}\omega) \quad \text{i.e.} \quad \psi_{\lambda_1}(t) = \lambda_1 \psi(\lambda_1 t). \tag{1}$$

The wavelet transform convolves a signal x with the wavelet filter bank. The scalogram x_1 localize the power spectrum of $x(t)$ around the log-frequencies $\log_2 \lambda_1 = j_1 + \frac{\chi_1}{Q}$ over durations $2Q\lambda_1^{-1}$, trading frequency resolution for time resolution:

$$x_1(t, \log \lambda_1) = |x \overset{t}{*} \psi_{\lambda_1}(t)| \quad \text{for all } \lambda_1 > 0\,, \tag{2}$$

$$x_1(t, \log_2 \lambda_1) = |x * \psi_{\lambda_1}|(t). \tag{3}$$

The constant-Q transform (CQT) $S_1 x$ corresponds to a low-pass filtering of x_1 with a window $\phi(t)$ of size T.

$$S_1 x(t, \log_2 \lambda_1) = x_1 * \phi_T(t) = |x * \psi_{\lambda_1}| * \phi_T(t). \tag{4}$$

There is a well-known dilemma in choosing T. Too small, the constant-Q matrix lacks invariance to time shifts, which will prevent any learning step to generalize from S_1x ; too large, discriminative information is discarded.

In order to combine the best of both worlds, the scattering transform recovers finer time scales than T with a second filterbank of wavelets $\psi_{\lambda_2}(t)$ of center frequencies λ_2, and applies complex modulus to improve regularity [6]. The wavelets $\psi_{\lambda_2}(t)$ have a quality factor in the range 1–2, though we choose to keep the same notation ψ for simplicity.

$$x_2(t, \log_2 \lambda_1, \log_2 \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}|\,(t) \quad (5)$$

Also known as *amplitude modulation spectrum*, the three-way array x_2 is then averaged in time to achieve as much invariance as the constant-Q spectrum S_1x:

$$S_2x(t, \log_2 \lambda_1, \log_2 \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}|\,(t) * \phi_T(t). \quad (6)$$

The concatenated scattering representation $Sx = \{S_1x, S_2x\}$ has proven to achieve higher accuracy in music genre classification as well as phoneme recognition [6] than audio features derived from S_1x only, such as Mel-frequency cepstral coefficients (MFCC).

2.2. Joint time-frequency scattering

Due to the constant-Q property, Sx is stable to small time warps of $x(t)$, as long as they do not exceed Q^{-1}, i.e. one semitone. This implies that small modulations, such as tremolo and vibrato, are accurately linearized in rate and depth [7].

However, the definition above is unstable to the variability in pitch and spectral envelope, for which the activations of frequency bands is highly correlated in time. To stabilize x_2 with respect to these variations, Andén [8] has redefined the wavelets ψ_{λ_2}'s as two-dimensional functions of both time and log-frequency, indexed by pairs $\lambda_2 = (\alpha, \beta)$, where α is measured in Hertz and β is measured in cycles per octaves.

$$\psi_{\lambda_2}(t, \log_2 \lambda_1) = \psi_\alpha(t) \times \psi_\beta(\log_2 \lambda_1) \quad (7)$$

The equation below introduces a "joint time-frequency scattering" transform, as opposed to the plain "time scattering" transform of Equation 5:

$$x_2(t, \log_2 \lambda_1, \log_2 \lambda_2) = |x_1 * \psi_{\lambda_2}(t, \log_2 \lambda_1)|. \quad (8)$$

The joint time-frequency scattering transform corresponds to the "cortical transform" introduced by Shamma to formalize his findings in auditory neuroscience.

2.3. Spiral scattering

The time-frequency scattering transform presented above provides template-free features for pitch variability along time. However, it is unaware of the harmonic structure in quasi-periodic signals, which are ubiquitous in audio recordings. The temporal evolution of this structure yields relevant information about attack transients and formantic changes, almost independently from the pitch contour.

In order to disentangle variabilities in pitch and spectral envelope, we extend the joint time-frequency scattering transform

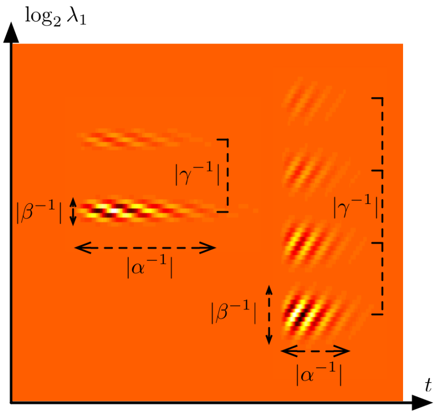

Figure 2: Two spiral wavelets $\psi_{\lambda_2}(t, \log_2 \lambda_1, \lfloor \log_2 \lambda_1 \rfloor)$ in the time-frequency plane, with different values of $\lambda_2 = (\alpha, \beta, \gamma)$. Left: $\alpha^{-1} = 120$ms, $\beta^{-1} = -0.25$ octave, $\gamma^{-1} = +2$ octaves. Right: $\alpha^{-1} = 60$ ms, $\beta^{-1} = +0.5$ octave, $\gamma^{-1} = -4$ octaves. Darker color levels corresponds to greater values of the real part.

to encompass motion across octaves, in conjunction with motion along neighboring constant-Q bands. We roll up the log-frequency variable $\log_2 \lambda_1$ into a Shepard pitch spiral (see Fig. 1), making one full turn at each octave. Since a frequency interval of one octave corresponds to one unit in binary logarithms $\log_2 \lambda_1$, pitch chroma and pitch height in the Shepard spiral correspond to integer part $\lfloor \log_2 \lambda_1 \rfloor$ and fractional part $\{\log_2 \lambda_1\}$:

$$\log_2 \lambda_1 = \lfloor \log_2 \lambda_1 \rfloor + \{\log_2 \lambda_1\}. \quad (9)$$

In this setting, the fundamental frequency f_0 is aligned with its power-of-two harmonics $2f_0$, $4f_0$, $8f_0$ and so forth. Likewise, the perfect fifth $3f_0$ is aligned with $6f_0$. As the number of harmonics per octave increase exponentially, the alignment of upper harmonics — $5f_0$, $7f_0$, and so forth — in the spiral is less crucial, because it can also be recovered with short-range time-frequency scattering.

We cascade three one-dimensional wavelet transforms in time, log-frequency, and octave index, to build a so-called Shepard spiral scattering transform, or alternatively "Shepardlet transform":

$$\psi_{\lambda_2}(t, \log_2 \lambda_1) = \psi_\alpha(t) \times \psi_\beta(\log_2 \lambda_1) \times \psi_\gamma(\lfloor \log_2 \lambda_1 \rfloor) \quad (10)$$

The definitions for x_2 and S_1x are the same as Equations **??** and **??**. Since its Fourier transform $\widehat{\psi_{\lambda_2}}$ is centered at (α, β, γ), the spiral wavelet ψ_{λ_2} has a pitch chroma velocity of α/β and a pitch height velocity of α/γ. Both velocities are measures in octaves per second.

Rolling up pitches into a spiral is a well-established idea in music, if only because of circularity of musical pitch classes. It has been studied by R. Shepard, J.C. Risset, and D. Deutsch to build paradoxes in perception of pitch, and has been validated by functional imaging of the auditory cortex[**?**].

3. DEFORMATIONS OF THE SOURCE-FILTER MODEL

Let $e(t) = \sum_n \delta(t - 2\pi f_0^{-1} n)$ be a harmonic signal and $t \mapsto \theta(t)$ a time warp function. We define a warped source as $e_\theta(t) = (e \circ$

$\theta)(t)$. Similarly, we compose a filter $h(t)$ and a warp $t \mapsto \nu(t)$ to define $h_\nu(t) = (h \circ \nu)(t)$. The warped source-filter model is

$$x(t) = [e_\theta * h_\nu](t). \tag{11}$$

Observe that $\dot{\theta}(t)$ induces a change of fundamental frequency, whereas $\dot{\nu}(t)$ accounts for a local dilation of the spectral envelope $|\widehat{h}|(\omega)$. We show in this section that, for $\dot{\theta}(t)$ and $\dot{\nu}(t)$ reasonably regular over the support of first-order wavelets, the local maxima of x_2 are clustered on a plane in the (α, β, γ) space of scattering coefficients. This plane satisfies the Cartesian equation

$$\alpha + \frac{\ddot{\theta}(t)}{\dot{\theta}(t)}\beta + \frac{\ddot{\nu}(t)}{\dot{\nu}(t)}\gamma = 0. \tag{12}$$

This result is likely to help automated transcription of polyphonic music, since notes overlapping both in time and frequency could be disentangled according to their respective source-filter velocities. Let $e_{\theta,1}(t, \log_2 \lambda_1)$ and $h_{\nu,1}(t, \log_2 \lambda_1)$ be the respective scalograms of $e_\theta(t)$ and $h_\nu(t)$. Our proof is driven by two properties:

1. *Harmonicity*. For any small octave difference $|j| \in \mathbb{N}$,

$$e_{\theta,1}(t, \log_2 \lambda_1) \approx e_{\theta,1}(t, \log_2 \lambda_1 + j). \tag{13}$$

2. *Spectral smoothness*. For any chroma difference $|\chi| < 1$,

$$h_{\nu,1}(t, \log_2 \lambda_1) \approx h_{\nu,1}(t, \log_2 \lambda_1 + \chi). \tag{14}$$

Given λ_1 near $pf_0\dot{\theta}(t)$ where $p \in \mathbb{N}$, the first step is to linearize $\theta(t)$ and $\nu(t)$ over the support of the first-order wavelet $\psi_{\lambda_1}(t)$. We work with the following assumptions:

(a) Q large enough to discriminate the p^{th} partial: $Q > 2p$,

(b) slowly varying source: $\|\ddot{\theta}/\dot{\theta}\|_\infty \ll \lambda_1/Q$, and

(c) slowly varying filter: $\|\ddot{\nu}/\dot{\nu}\|_\infty \ll \lambda_1/Q$.

According to (a), partials $p' \neq p$ have a negligible contribution to e_1 at the log-frequency $\log_2 \lambda_1$. For lack of any interference, e_1 is constant through time, and we may drop the dependency in t:

$$e_1(\log_2 \lambda_1) = |\widehat{\psi_{\lambda_1}}(pf_0)| \tag{15}$$

According to (b), the scalogram of the deformed source can be replaced by the scalogram of the original source translated along the log-frequency at the velocity $\log_2 \dot{\theta}(t)$:

$$e_{\theta,1}(t, \log_2 \lambda_1) = e_1(\log_2 \lambda_1 - \log_2 \dot{\theta}(t)). \tag{16}$$

Similarly, we leverage (c) to linearize $\nu(t)$ over the support of $\psi_{\lambda_1}(t)$. The spectral smoothness assumption allows to approximate $\hat{h}(\omega)$ by a constant over the frequential support of the wavelet, hence to factorize the filtering as a product:

$$[h_\nu * \psi_{\lambda_1}] = h_1(\log_2 \lambda_1 - \log_2 \dot{\nu}(t))\psi_{\lambda_1}\left(\frac{\nu(t)}{\dot{\nu}(t)}\right). \tag{17}$$

By plugging Equation 16 into Equation 17, x_1 appears is a separable product between e_1 and h_1, moving in log-frequency at respective velocities $\log_2 \dot{\theta}(t)$ and $\log_2 \dot{\nu}(t)$:

$$x_1(t, \log_2 \lambda_1) = e_1(\log_2 \lambda_1 - \log_2 \dot{\theta}(t))h_1(\log_2 \lambda_1 - \log_2 \dot{\nu}(t)). \tag{18}$$

The second step in the proof consists in showing that the convolution along chromas with ψ_β only applies to $e_{1,\theta}$, whereas the convolution across octaves with ψ_γ only applies to $h_{1,\nu}$. Indeed, all wavelets are designed to carry a negligible mean value, i.e. convolving them with a constant yields zero. Therefore, the harmonicity and spectral smoothness properties rewrite as

$$e_{\theta,1} \overset{j}{*} \psi_\gamma \approx 0 \quad \text{and} \quad h_{\nu,1} \overset{\chi}{*} \psi_\beta \approx 0. \tag{19}$$

Gathering Equations 18 and 19 into the definition of spiral scattering yields

$$x_1 \overset{t,\chi,j}{*} \psi_{\lambda_2} = \left[\left(e_{1,\theta} \overset{\chi}{*} \psi_\beta\right) \times \left(h_{1,\nu} \overset{j}{*} \psi_\gamma\right)\right] \overset{t}{*} \psi_\alpha, \tag{20}$$

where the superscripts t, χ and j denote convolutions along time, chromas and octaves respectively.

As a final step, we state that the phase of $[e_{\theta,1} \overset{\chi}{*} \psi_\beta]$ is $\beta \times (\log_2 \lambda_1 - \log_2 p\dot{\theta}(t))$. By differentiating this quantity along t for fixed $\log_2 \lambda_1$, we obtain an instantaneous frequency of $-\beta\ddot{\theta}(t)/\dot{\theta}(t)$. Similarly, the instantaneous frequency of the convolution $[h_{\nu,1} \overset{j}{*} \psi_\gamma]$ is $-\gamma\ddot{\nu}(t)/\dot{\nu}(t)$. As long as

$$\alpha \geq \left|\frac{\ddot{\theta}(t)}{\dot{\theta}(t)}\beta\right| \quad \text{and} \quad \alpha \geq \left|\frac{\ddot{\nu}(t)}{\dot{\nu}(t)}\gamma\right|, \tag{21}$$

the envelopes of these two convolutions are almost constant over the support of $\psi_\alpha(t)$. We conclude with the following approximate closed-form expression for the spiral scattering coefficients of the deformed source-filter model:

$$x_2(t, \log_2 \lambda_1, \log_2 \lambda_2) = |e_{\theta,1} \overset{\chi}{*} \psi_\beta| \times |h_{\nu,1} \overset{j}{*} \psi_\gamma| \times \left|\widehat{\psi_\alpha}\left(-\frac{\ddot{\theta}(t)}{\dot{\theta}(t)}\beta - \frac{\ddot{\nu}(t)}{\dot{\nu}(t)}\gamma\right)\right|. \tag{22}$$

The Fourier spectrum $|\widehat{\psi_\alpha}(\omega)|$ of $\psi_\alpha(t)$ is a bump centered at the frequency α. Equation 12 follows immediately from the above formula. The same result holds for the averaged coefficients $S_2 x = x_2 * \phi_T(t)$ if

$$\left|\frac{\dddot{\theta}(t)}{\ddot{\theta}(t)} - \frac{\ddot{\theta}(t)}{\dot{\theta}(t)}\right| \ll T^{-1} \quad \text{and} \quad \left|\frac{\dddot{\nu}(t)}{\ddot{\nu}(t)} - \frac{\ddot{\nu}(t)}{\dot{\nu}(t)}\right| \ll T^{-1}. \tag{23}$$

An important caveat is that the inequalities above do not hold at inflexion points of the diffeomorphisms $\theta(t)$ and $\nu(t)$, i.e. where the relative velocities $\ddot{\theta}(t)/\dot{\theta}(t)$ or $\ddot{\nu}(t)/\dot{\nu}(t)$ cross zero.

4. CONCLUSIONS

The spiral model we have presented is well-known in music theory and experimental psychology. However, existing methods in audio signal processing do not fully take advantage from its richness, as they either picture pitch on a line (e.g. MFCC) or on a circle (e.g. chroma features).

Future work will be devoted to evaluating the discriminative power of Shepard spiral scattering coefficients over a variety of classification pipelines. Our representation also encompass automatic music transcription, perceptual similarity learning, and new audio transformations as potential applications.

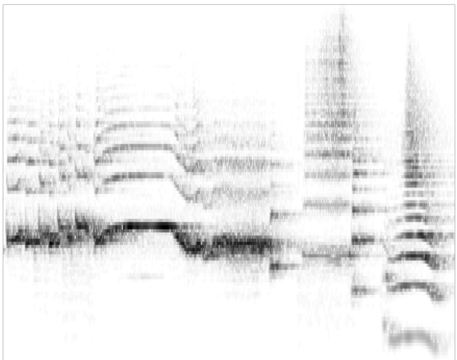

Figure 3

5. REFERENCES

[1] S. Abdallah and M. Plumbley, “Polyphonic music transcription by non-negative sparse coding of power spectra,” in *Proc. ISMIR*, 2004, vol. 510, pp. 10–14.

[2] P. Smaragdis and J.C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, 2003.

[3] R. Hennequin, R. Badeau, and B. David, “NMF with time-frequency activations to model nonstationary audio events,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744–753, 2011.

[4] B. Fuentes, R. Badeau, and G. Richard, “Harmonic adaptive latent component analysis of audio and application to music transcription,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 9, pp. 1854–1866, 2013.

[5] R. Shepard, “Circularity in Judgments of Relative Pitch,” 1964.

[6] J. Andén and S. Mallat, “Deep Scattering Spectrum,” *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[7] Joakim Andén and Stéphane Mallat, “Scattering representation of modulated sounds,” *Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx-12)*, , no. 3, pp. 15–18, 2012.

[8] Joakim Andén, *Time and frequency scattering for audio classification*, Ph.D. thesis, École Polytechnique, 2014.