ABSTRACT

We present a new representation of harmonic sounds that linearizes the dynamics of pitch and spectral envelope, 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 pitch spiral, in that convolutions are successively performed in time, log-frequency, and octave index. We give a closed-form approximation of spiral scattering coefficients for a nonstationary generalization of the harmonic source-filter model.

1. INTRODUCTION

The spectro-temporal evolution of harmonic spectra conveys essential information to audio classification, blind source separation, transcription, as well as other processing tasks. This information is however difficult to capture in time-varying, polyphonic mixtures. On one hand, spectrogram-based pattern recognition algorithms [1] are exposed to detection errors as they impose strong constraints on the shape of harmonic templates. On the other, time-varying generalizations of matrix factorization [2] are under-constrained and thus may fail to converge to a satisfying solution. In this article, we address the characterization of harmonic structures without any detection nor training step.

Wavelets have long proven to provide meaningful, sparse activations as long as they operate on a dimension on which the signal has already some regularity. Although a single sine wave draws a regular edge on the time-frequency plane, a harmonic comb is made of distant sharp peaks over the log-frequency axis, an irregular pattern that is hard to characterize globally. This irregularity weakens the discriminative power of existing wavelet-like representations, such as Mel-frequency cepstral coefficients (MFCC).

To recover regularity across partials within a wavelet framework, we capitalize on the fact that power-of-two harmonics are exactly one octave apart. 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 time and log-frequency domain.

Once specified the variables of time, log-frequency, and octave index, our representation merely consists in cascading three wavelet decompositions along them and applying complex modulus. Thus, the constant-Q scalogram is "scattered" into channels over which main factors of time variability are disentangled and regularized, yet harmonicity is preserved.

---

Vincent Lostanlen, Stéphane Mallat*

Department of Computer Science, École normale supérieure

Paris, France

vincent.lostanlen@ens.fr

*This work is supported by the ERC InvariantClass 320959. The source code to reproduce figures and experiments is available at www.github.com/lostanlen/scattering.m.
The wavelet transform convolves an input signal \( x(t) \) with the filter bank of \( \psi_{\lambda} \)’s. We denote convolutions along time by the operator \( \star \). Applying complex modulus to all wavelet convolutions results in the "scalogram" matrix

\[
x_1(t, \log \lambda_1) = |x \star \psi_{\lambda_1}| \quad \text{for all } \lambda_1 > 0,
\]

whose principal axis is uniformly sampled by the binary logarithm \( \log \lambda_1 \). The scalogram \( x_1 \) localizes the energy of \( x(t) \) around frequencies \( \lambda_1 > 0 \) over durations \( 2Q \lambda_1^{-1} \), trading frequency resolution for time resolution.

The constant-Q transform (CQT) \( S_1 \) corresponds to the fast implementation of wavelet scattering, yielding

\[
S_1 x(t, \log \lambda_1) = x_1 \star \phi_T = |x \star \psi_{\lambda_1} | \star \phi_T.
\]

2.2. Joint time-frequency scattering

The time scattering transform defined in Equation (4) decomposes \( x_2 \) into a second filterbank of wavelets \( \psi_{\lambda_2} \) and applies complex modulus to get

\[
x_2(t, \log \lambda_1, \log \lambda_2) = |x \star \psi_{\lambda_2}| = | |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|.
\]

The wavelets \( \psi_{\lambda_2} \) have a quality factor in the range 1–2, though we choose to keep the same notation \( \psi \) for simplicity. Like in Equation (3), averaging in time creates invariance to translation in time up to \( T \), yielding

\[
S_2 x(t, \log \lambda_1, \log \lambda_2) = x_2 \star \phi = | |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_T.
\]

Due to the constant-Q property, \( S_1 x \) and \( S_2 x \) are 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 [3][4].

2.3. Spiral scattering

The time-frequency scattering transform defined in Equation (7) provides template-free features for pitch variability along time. However, it is unaware of the harmonic structure of voiced sounds, such as vowels or musical notes. The temporal evolution of this structure yields relevant information about attack transients and formantic changes, almost independently from the pitch contour.

In order to capture this information, we extend the joint time-frequency scattering transform to encompass regularity in time across octaves of fixed chroma, in conjunction with regularity along neighboring constant-Q bands. Just like wavelet filterbanks along time and log-frequency have been defined in the two previous subsections, we deal here with regularity across octaves by introducing a third wavelet filterbank.

We roll up the log-frequency variable \( \log \lambda_1 \) into a pitch spiral making one full turn at each octave (see Figure 2). Since a frequency interval of one octave corresponds to one unit in binary logarithm \( \log \lambda_1 \), pitch height and pitch chroma in the spiral correspond to integer part \( \lfloor \log \lambda_1 \rfloor \) and fractional part \( \{ \log \lambda_1 \} \):

\[
\log \lambda_1 = \lfloor \log \lambda_1 \rfloor + \{ \log \lambda_1 \} = j_1 + \frac{\lambda_1}{Q}.
\]

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 convolutions along chromas for \( \beta^{-1} \) — of the order of a few semitones.

The wavelet \( \psi_{\lambda_2} \) is now defined as a product between wavelets \( \psi_\) along in time, log-frequency, and octave index:

\[
\psi_{\lambda_2}(t, \log \lambda_1) = \psi_\alpha(t) \times \psi_\beta(\log \lambda_1) \times \psi_\gamma(\{ \log \lambda_1 \}).
\]
Examples of the "spiral wavelet" $\psi_{32}$ are shown in Figure 2 for different values of $\alpha, \beta$ and $\gamma$. To ensure invertibility and energy conservation, the quencies $\beta$ and $\gamma$ must take negative values, including zero. We adopt the shorthand notation

$$\log \lambda = (\log \alpha, \log |\beta|, \log \gamma, \log \gamma)$$

(10)

for their indexing. In the special case $\beta = 0$, $\psi_{32}$ is no longer a wavelet but a low-pass filter whose support covers one octave. By convention, the corresponding log-frequency index is $\log |\beta| = -\infty$. The same remark applies to $\psi_{32}$ for $\gamma = 0$, which covers six octaves. Since its Fourier transform $\hat{\psi}_{32}$ is centered at $(\alpha, \beta, \gamma)$, the spiral wavelet $\psi_{32}$ has a pitch chroma velocity of $\alpha \beta$ and a pitch height velocity of $\alpha \gamma$, both measured in octaves per second.

We write $\dagger$ to denote convolutions across neighboring octaves. The definition for $x_2$ is comparable to Equations (1) and (3):

$$x_2(t, \log \lambda_1, \log \lambda_2) = |x_1^{\gamma_1} \psi_{31} + \psi_{32}|$$

(11)

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 Shepard [6], Risset [7], and Deutsch [8] to build paradoxes in perception of pitch, and is corroborated by functional imaging of the auditory cortex [9].

3. DEFORMATIONS OF THE SOURCE-FILTER MODEL

A classical model for voiced speech production consists in the convolution of a harmonic glottal source $e(t)$ with a vocal tract filter $h(t)$. Introducing independent deformations to both components brings realistic variability to pitch and spectral envelope. This section studies the decomposition of the deformed source-filter model in the spiral scattering transform.

3.1. Overview

Let $e(t) = \sum_{n} \delta(t - 2\pi f_0 n)$ be a harmonic signal and $h(t)$ a warping function. We define a warped source as $e_{\nu}(t) = (e \circ h)(t)$. Similarly, we compose a filter $h_{\nu}(t)$ and a warping $\nu(t)$ to define $h_{\nu}(t) = (h \circ \nu)(t)$. The warped source-filter model is

$$x(t) = [e_{\nu} \times h_{\nu}](t).$$

(12)

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

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

(13)

This result means that harmonic sounds overlapping both in time and frequency could be resolved according to their respective source-filter velocities.

Our proof is driven by harmonic and spectral smoothness properties — Equation (13) — and derives Equation (13) from the computation of wavelet ridges on the pitch spiral [10].

3.2. Source-filter factorization in the scalogram

Given $\lambda_1 \approx p \dot{\theta}(t)$ where $p \in \mathbb{N}$, we linearize $\dot{\theta}(t)$ and $\nu(t)$ over the support of the first-order wavelet $\psi_{31}(t)$. We work under the following assumptions:

(a) $Q$ large enough to discriminate the $p$th partial: $Q > 2p$,
(b) slowly varying source: $|\dot{\theta}/\dot{\theta}| \ll \lambda_1/Q$,
(c) slowly varying filter: $|\dot{\nu}/\dot{\nu}| \ll \lambda_1/Q$, and
(d) spectral smoothness: $|\dot{h}/\dot{h}| \ll Q/\lambda_1$.

According to (a), partials $p' \neq p$ have a negligible contribution to the scalogram $e_1 = [x_{\nu} \times \psi_{31}](t)$ of $e(t)$ at the log-frequency $\log \lambda_1$. For lack of any interference, $e_1$ is constant through time, and we may drop the dependency in $t$:

$$e_1(\log \lambda_1) = [\ddot{\psi}_{31}(p f_0)].$$

(14)

According to (b), the scalogram $e_{\nu,1} = [e_{\nu} \times \psi_{31}](t)$ of the deformed source $e_{\nu}(t)$ can be replaced by the scalogram of the original source translated along the log-frequency at the velocity $\log \dot{\theta}(t)$:

$$e_{\nu,1}(t, \log \lambda_1) = e_1(\log \lambda_1 - \log \dot{\theta}(t)) + \nu(t).$$

(15)

According to (c), we linearize $\nu(t)$ over the support of $\psi_{31}(t)$.

According to (d), we approximate $\dot{h}(\omega)$ by a constant over the fundamental support of the wavelet and factorize the filtering as a product:

$$[h_{\nu} \times \psi_{31}] = h_1(\log \lambda_1 - \log \dot{\nu}(t)) \times \psi_{31}(t) \frac{\dot{\nu}(t)}{\dot{\theta}(t)}.$$

(16)

By plugging Equation (15) into Equation (10), $x_1$ appears as a separable product between $e_1$ and $h_1$, moving in log-frequency at respective velocities $\log \dot{\theta}(t)$ and $\log \dot{\nu}(t)$:

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

(17)

3.3. Harmonicity and spectral smoothness properties

The second step in the proof consists in showing that the convolution along chromas with $\psi_{31}$ only applies to $e_{1,\nu}$, whereas the convolution across octaves with $\psi_3$ 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

$$e_{\nu,1} \times \psi_{3} \approx 0 \quad \text{and} \quad h_{1,\nu} \times \psi_{31} \approx 0.$$

(18)

Gathering Equations (17) and (18) into the definition of spiral scattering yields

$$x_1^{\gamma_1, j_1} \times \psi_{32} = e_{1,\nu} \times \psi_{3} \times h_{1,\nu} \times \nu \psi_{31} \times \psi_{3} \times \psi_{3} \times \psi_{3} \times \psi_{3}.$$

(19)
3.4. Extraction of instantaneous frequencies

As a final step, we state that the phase of \( [e_{\theta_1} \times \psi_1] \) is \( \beta \times (\log \lambda_1 - \log \beta(t)) \). By differentiating this quantity along \( \beta \) for fixed \( \log \lambda_1 \), we obtain an instantaneous frequency of \( -\beta(t)/\beta(t) \). Similarly, the instantaneous frequency of the filter after convolution across octaves \( [h_{\nu_1} \times \nu_1] \) is \( -\gamma(t)/\nu(t) \). As long as

\[
\alpha \geq \frac{\dot{\theta}(t)}{\dot{\theta}(t)} \beta \quad \text{and} \quad \alpha \geq \frac{\dot{\nu}(t)}{\nu(t)} \gamma ,
\tag{20}
\]

the envelopes of these two convolutions are almost constant over the support of \( \psi_1(t) \). We conclude with the following approximate closed-form expression for the spiral scattering coefficients of the deformed source-filter model:

\[
x_{2}(t, \log \lambda_1, \log \lambda_2) = [e_{\theta_1} \times \psi_1] \times [h_{\nu_1} \times \nu_1] \times \beta_{\alpha} \frac{\dot{\theta}(t)}{\dot{\theta}(t)} \beta - \frac{\dot{\nu}(t)}{\nu(t)} \gamma .
\tag{21}
\]

The Fourier spectrum \( |\beta_{\alpha}(\omega)| \) of \( \psi_1(t) \) is a bump centered at the frequency \( \alpha \). Equation 13 follows immediately from the above formula. The same result holds for the averaged coefficients \( S_{2} \hat{x} = x_{2} \circ \phi(t) \) if the velocities \( \theta(t)/\dot{\theta}(t) \) or \( \dot{\nu}(t)/\nu(t) \) have small relative variations:

\[
\frac{\dot{\theta}(t)}{\dot{\theta}(t)} - \frac{\dot{\theta}(t)}{\dot{\theta}(t)} \ll T^{-1} \quad \text{and} \quad \frac{\dot{\nu}(t)}{\nu(t)} - \frac{\dot{\nu}(t)}{\nu(t)} \ll T^{-1} .
\tag{22}
\]

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

In the example of a trombone signal, glissando can be modeled by \( \dot{\theta}(t)/\dot{\theta}(t) \) in the source-filter model, whereas the brassiness profile induces a timbral velocity \( \dot{\nu}(t)/\nu(t) \). Figure 3 illustrates that these two velocities are stably disentangled and characterized.

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). In this article, we have shown how spiral scattering can represent the transientness of harmonic sounds.

5. REFERENCES


