WAVELET SCATTERING ON THE PITCH SPIRAL

*Vincent Lostanlen, Stéphane Mallat**

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

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 enforce strong constraints on the shape of harmonic templates. On the other, time-varying generalizations of matrix factorization [2] are underconstrained 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 into 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 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.

* 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.

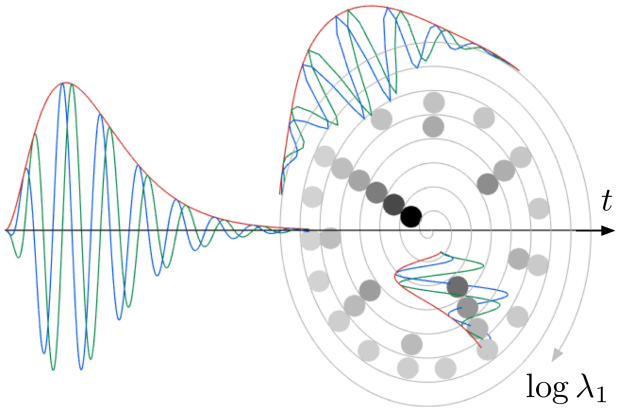

Figure 1: The spiral wavelet is a product of wavelets along time, log-frequency, and octave index. Blue and green oscillations represent the real and imaginary parts. The red envelope represents the complex modulus. Partials of an hypothetical harmonic sound are marked as thick dots.

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 scattering coefficients of a nonstationary musical note.

2. FROM TIME SCATTERING TO SPIRAL SCATTERING

This section builds the spiral scattering transform progressively as a cascade of wavelet transforms along time, log-frequency, and octave index. All three variables share the same framework.

2.1. Time scattering

An analytic "mother" wavelet is a complex filter $\psi(t)$ whose Fourier transform $\widehat{\psi}(\omega)$ is concentrated over the dimensionless frequency interval $[1 - 1/2Q, 1 + 1/2Q]$, where the quality factor Q is in the typical range 12–24. Dilations of this wavelet define a family of bandpass filters centered at frequencies $\lambda_1 = 2^{j_1 + \frac{\chi_1}{Q}}$, where the indices $j_1 \in \mathbb{Z}$ and $\chi_1 \in \{1 \dots Q\}$ respectively denote octave and chroma:

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

The wavelet transform convolves an input signal $x(t)$ with the filter bank of ψ_{λ_1}'s. We denote convolutions along time by the operator $\overset{t}{*}$. Applying complex modulus to all wavelet convolutions results in the "scalogram" matrix

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

whose frequential axis is uniformly sampled by the binary logarithm $\log \lambda_1$. The scalogram x_1 localizes the energy of $x(t)$ around frequencies λ_1 over durations $2Q/\lambda_1$, trading frequency resolution for time resolution.

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

$$S_1 x(t, \log \lambda_1) = x_1 \overset{t}{*} \phi_T = |x \overset{t}{*} \psi_{\lambda_1}| \overset{t}{*} \phi_T. \tag{3}$$

To recover the amplitude modulations lost when averaging by ϕ_T in Equation (3), the time scattering transform also convolves x_1 with a second filterbank of wavelets ψ_{λ_2} and applies complex modulus to get

$$x_2(t, \log \lambda_1, \log \lambda_2) = |x_1 \overset{t}{*} \psi_{\lambda_2}| = ||x \overset{t}{*} \psi_{\lambda_1}| \overset{t}{*} \psi_{\lambda_2}|. \tag{4}$$

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. Like in Equation (3), averaging in time creates invariance to translation in time up to T, yielding

$$S_2 x(t, \log \lambda_1, \log \lambda_2) = x_2 \overset{t}{*} \phi_T = ||x \overset{t}{*} \psi_{\lambda_1}| \overset{t}{*} \psi_{\lambda_2}| \overset{t}{*} \phi_T. \tag{5}$$

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].

2.2. Joint time-frequency scattering

The time scattering transform defined in Equation (4) decomposes each frequency band separately, and so cannot properly capture the coherence of time-frequency structures, such as those induced by pitch contour. To remedy this, Andén [5] has redefined the wavelets ψ_{λ_2}'s as functions of both time and log-frequency, indexed by pairs $\lambda_2 = (\alpha, \beta)$, where α is a modulation frequency in Hertz and β is a "quefrency" in cycles per octaves. The joint wavelets $\psi_{\lambda_2}(t, \log \lambda_1)$ factorize as

$$\psi_{\lambda_2}(t, \log \lambda_1) = \psi_\alpha(t) \times \psi_\beta(\log \lambda_1). \tag{6}$$

We write $\overset{\chi_1}{*}$ to denote convolutions along the log-frequency axis, i.e. along chromas. Wavelet scattering is extended to two-dimensional convolutions by plugging Equation (6) into the definition of x_2 in Equation (4):

$$x_2(t, \log \lambda_1, \log \lambda_2) = |x_1 \overset{t,\chi_1}{*} \psi_{\lambda_2}| = |x_1 \overset{t}{*} \psi_\alpha \overset{\chi_1}{*} \psi_\beta|. \tag{7}$$

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

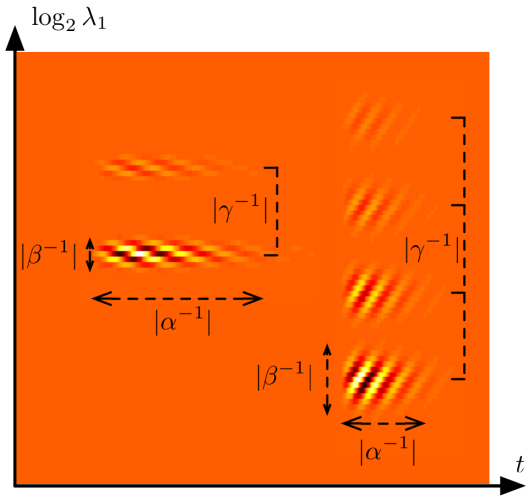

Figure 2: Two spiral wavelets $\psi_{\lambda_2}(t, \log \lambda_1)$ in the time-frequency plane, with different values of $\lambda_2 = (\alpha, \beta, \gamma)$. Left: $\alpha^{-1} = 120$ ms, $\beta^{-1} = -0.25$ octave, $\gamma^{-1} = +2$ octaves. Right: $\alpha^{-1} = 60$ ms, $\beta^{-1} = +0.5$ octave,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

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 at 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 capitalize on harmonicity 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 1). Since a frequency interval of one octave corresponds to one unit in binary logarithms $\log \lambda_1$, pitch height and pitch chroma in the spiral correspond to integer part $\lfloor \log \lambda_1 \rfloor$ and fractional part $\{\log \lambda_1\}$:

$$\log \lambda_1 = \lfloor \log \lambda_1 \rfloor + \{\log \lambda_1\} = j_1 + \frac{\chi_1}{Q}. \tag{8}$$

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 β^{-1} of the order of a few semitones.

The wavelet ψ_{λ_2} is now defined as a product between wavelets in time, log-frequency, and octave index:

$$\psi_{\lambda_2}(t, \log \lambda_1) = \psi_\alpha(t) \times \psi_\beta(\log \lambda_1) \times \psi_\gamma(\lfloor \log \lambda_1 \rfloor). \tag{9}$$

Examples of the "spiral wavelet" ψ_{λ_2} are shown in Figure 2 for different values of α, β and γ. To ensure invertibility and energy conservation, the quefrencies β and γ must take negative values, including zero. We adopt the shorthand notation

$$\log \lambda_2 = (\log \alpha, \log |\beta|, \operatorname{sign} \beta, \log |\gamma|, \operatorname{sign} \gamma) \tag{10}$$

to specify their indexing. In the special case $\beta = 0$, ψ_β is no longer a wavelet but a low-pass filter whose support covers one octave. By convention, the corresponding log-quefrency index is $\log |\beta| = -\infty$. The same remark applies to ψ_γ for $\gamma = 0$, which covers six octaves. 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 measured in octaves per second.

We write $\overset{j_1}{*}$ to denote convolutions across neighboring octaves. The definition for x_2 is comparable to Equations (4) and (7):

$$\begin{aligned} x_2(t, \log \lambda_1, \log \lambda_2) &= |x_1 \overset{t,\chi_1,j_1}{*} \psi_{\lambda_2}| \\ &= |x_1 \overset{t}{*} \psi_\alpha \overset{\chi_1}{*} \psi_\beta \overset{j_1}{*} \psi_\gamma|. \end{aligned} \tag{11}$$

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 $e(t) = \sum_n \delta(t-n)$ 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 the signal

$$x_{\theta,\nu}(t) = (e_\theta \overset{t}{*} h_\nu)(t) \tag{12}$$

Observe that $\dot\theta(t)$ induces a change of fundamental frequency, whereas $\dot\nu(t)$ accounts for a local dilation of the spectral envelope $|\widehat{h}|(\omega)$. 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

$$\alpha + \frac{\ddot\theta(t)}{\dot\theta(t)}\beta + \frac{\ddot\nu(t)}{\dot\nu(t)}\gamma = 0. \tag{13}$$

In a polyphonic context, 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 harmonicity and spectral smoothness properties — Equation (18) — and derives Equation (13) from the computation of wavelet ridges on the pitch spiral [10].

3.2. Source-filter factorization in the scalogram

Given λ_1 near the p^{th} partial $p\dot\theta(t)$ where $p \in \mathbb{N}$, we linearize $\theta(t)$ and $\nu(t)$ over the support of the first-order wavelet $\psi_{\lambda_1}(t)$. We work under the following assumptions:

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

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

(c) slowly varying filter: $\|\ddot\nu/\dot\nu\|_\infty \ll \lambda_1/Q$, and

(d) spectral smoothness: $\|\mathrm{d}(\log|\hat h|)/\mathrm{d}\omega\|_\infty \times \|1/\dot\eta\| \ll Q/\lambda_1$.

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

$$|e \overset{t}{*} \psi_{\lambda_1}| \approx |\widehat{\psi}_{\lambda_1}(p)|. \tag{14}$$

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

$$|e_\theta \overset{t}{*} \psi_{\lambda_1}|(t) = |e \overset{t}{*} \psi_{\lambda_1}|(\theta(t)) \approx |\widehat{\psi}_{\lambda_1}(p\dot\theta(t))|. \tag{15}$$

According to (c), we linearize $\nu(t)$ over the support of $\psi_{\lambda_1}(t)$. According to (d), we approximate $\hat h(\omega)$ by a constant over the frequential support of the wavelet and factorize the filtering as a product:

$$(h_\nu * \psi_{\lambda_1})(t) \approx h_1(\log \lambda_1 - \log \dot\nu(t)) \times \psi_{\lambda_1}\left(\frac{\nu(t)}{\dot\nu(t)}\right). \tag{16}$$

By plugging Equation (15) into Equation (16), 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 $\log \dot\nu(t)$:

$$\begin{aligned} &x_1(t, \log \lambda_1) = \\ &\qquad e_1(\log \lambda_1 - \log \dot\theta(t)) h_1(\log \lambda_1 - \log \dot\nu(t)). \end{aligned} \tag{17}$$

3.3. Harmonicity and spectral smoothness properties

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 smothness properties rewrite as

$$e_{\theta,1} \overset{j_1}{*} \psi_\gamma \approx 0 \quad \text{and} \quad h_{\nu,1} \overset{\chi_1}{*} \psi_\beta \approx 0. \tag{18}$$

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

$$x_1 \overset{t,\chi_1,j_1}{*} \psi_{\lambda_2} = \left[\left(e_{1,\theta} \overset{\chi_1}{*} \psi_\beta\right) \times \left(h_{1,\nu} \overset{j_1}{*} \psi_\gamma\right)\right] \overset{t}{*} \psi_\alpha, \tag{19}$$

where the superscripts t, χ_1, and j_1 denote convolutions along time, chromas and octaves respectively.

3.4. Extraction of instantaneous frequencies

As a final step, we state that the phase of $(e_{\theta,1} \overset{\chi}{*} \psi_\beta)$ is $\beta \times (\log \lambda_1 - \log p\dot{\theta}(t))$. By differentiating this quantity along t for fixed $\log \lambda_1$, we obtain an instantaneous frequency of $-\beta\ddot{\theta}(t)/\dot{\theta}(t)$. Similarly, the instantaneous frequency of the filter scalogram after convolution across octaves $(h_{\nu,1} \overset{j_1}{*} \psi_\gamma)$ is $-\gamma\ddot{\nu}(t)/\dot{\nu}(t)$. As long as

$$\alpha \geq \left|\frac{\ddot{\theta}(t)}{\dot{\theta}(t)}\beta\right| \quad \text{and} \quad \alpha \geq \left|\frac{\ddot{\nu}(t)}{\dot{\nu}(t)}\gamma\right|, \tag{20}$$

the envelopes of these two convolutions are almost constant over the support of $\psi_\alpha(t)$ [10]. We conclude with the following approximate closed-form expression for the spiral scattering coefficients of the deformed source-filter model:

$$x_2(t, \log\lambda_1, \log\lambda_2) \approx$$
$$|e_{\theta,1} \overset{\chi}{*} \psi_\beta| \times |h_{\nu,1} \overset{j_1}{*} \psi_\gamma| \times \left|\widehat{\psi}_\alpha\left(-\frac{\ddot{\theta}(t)}{\dot{\theta}(t)}\beta - \frac{\ddot{\nu}(t)}{\dot{\nu}(t)}\gamma\right)\right|. \tag{21}$$

The Fourier spectrum $|\widehat{\psi}_\alpha(\omega)|$ of $\psi_\alpha(t)$ is a bump centered at the frequency α. Equation (13) follows immediately from the above formula. The same result holds for the averaged coefficients $S_2x = x_2 * \phi_T(t)$ if the velocities $\ddot{\theta}(t)/\dot{\theta}(t)$ or $\ddot{\nu}(t)/\dot{\nu}(t)$ have small relative variations:

$$\left|\frac{\dddot{\theta}(t)}{\ddot{\theta}(t)} - \frac{\ddot{\theta}(t)}{\dot{\theta}(t)}\right| \ll T^{-1} \quad \text{and} \quad \left|\frac{\dddot{\nu}(t)}{\ddot{\nu}(t)} - \frac{\ddot{\nu}(t)}{\dot{\nu}(t)}\right| \ll T^{-1}. \tag{22}$$

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 velocities $\ddot{\theta}(t)/\dot{\theta}(t)$ or $\ddot{\nu}(t)/\dot{\nu}(t)$ cross zero.

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

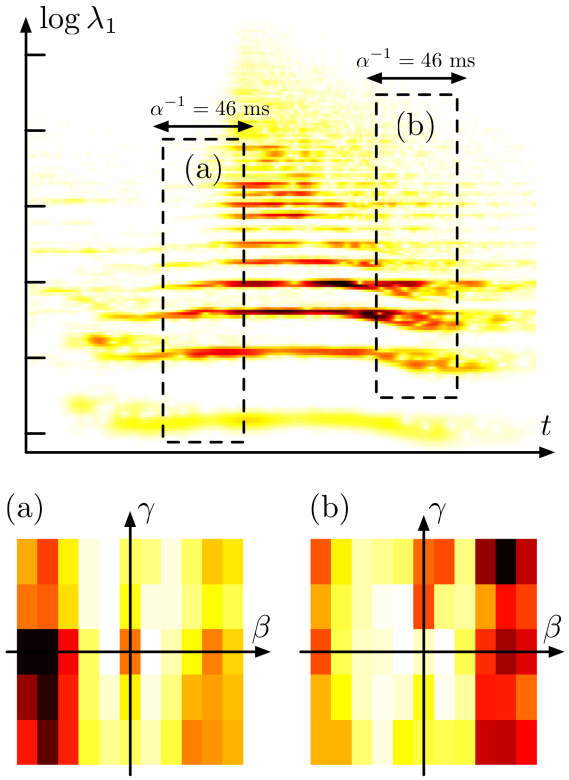

Figure 3: Top: scalogram of a musical note from Berio's *Sequenza V* for trombone, around 3'45". Observe that the attack part (a) has increasing pitch and increasing brightness, whereas the release part (b) has decreasing pitch and decreasing brightness. Bottom: spiral scattering coefficients for t and $\log\lambda_1$ specified by (a) and (b), α^{-1} fixed at 46 ms, β^{-1} ranging from -1 octave to $+1$ octave, and γ^{-1} ranging from -4 octaves to $+1$ octaves. As expected, highest values are concentrated in the bottom left corner for (a) and in the top right corner for (b).

4. CONCLUSIONS

The spiral model we have presented is well-known in music theory and experimental psychology [6, 7, 8]. However, existing methods in audio signal processing do not fully take advantage from its richness, because 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

[1] C. Kereliuk and P. Depalle, "Improved Hidden Markov Model Partial Tracking Through Time-frequency Analysis," in *Proc. DAFx*, 2008, pp. 1–6.

[2] R. Hennequin, R. Badeau, and B. David, "NMF with Time-frequency Activations to Model Nonstationary Audio Events," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 744–753, 2011.

[3] J. Andén and S. Mallat, "Scattering Representation of Modulated Sounds," *Proc. DAFx*, , no. 3, pp. 15–18, 2012.

[4] J. Andén and S. Mallat, "Deep Scattering Spectrum," *IEEE Trans. Sig. Proc.*, vol. 62, no. 16, pp. 4114–4128, 2014.

[5] J. Andén, *Time and Frequency Scattering for Audio Classification*, Ph.D. thesis, École Polytechnique, 2014.

[6] R. Shepard, "Circularity in Judgments of Relative Pitch," *J. Acoust. Soc. Am.*, vol. 36, no. 12, pp. 2346, 1964.

[7] J.-C. Risset, "Pitch control and pitch paradoxes demonstrated with computer-synthesized sounds," *J. Acoust. Soc. Am.*, vol. 46, no. 1A, pp. 88–88, 1969.

[8] D. Deutsch, K. Dooley, and T. Henthorn, "Pitch circularity from tones comprising full harmonic series.," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 589–597, 2008.

[9] J. Warren, S. Uppenkamp, R. Patterson, and T. Griffiths, "Analyzing Pitch Chroma and Pitch Height in the Human Brain," *Ann. N. Y. Acad. Sci.*, vol. 999, no. 17, pp. 212–214, 2003.

[10] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, and B. Torrésani, "Asymptotic Wavelet and Gabor Analysis: Extraction of Instantaneous Frequencies," *IEEE Trans. Inf. Theory*, vol. 38, pp. 644–664, 1992.