THE SHAPE OF REMIXXXES TO COME: AUDIO TEXTURE SYNTHESIS WITH TIME-FREQUENCY SCATTERING

Vincent Lostanlen

Music and Audio Research Lab
Steinhardt School of Culture, Education, and Human Development
New York University
New York, NY, USA
vincent.lostanlen@nyu.edu

Florian Hecker

Edinburgh College of Art
College of Arts, Humanities, and Social Sciences
University of Edinburgh
Edinburgh, UK
florian.hecker@ed.ac.uk

ABSTRACT

This article explains how to apply time–frequency scattering, a convolutional operator extracting modulations in the time–frequency domain at different rates and scales, to the re-synthesis and manipulation of audio textures. After implementing phase retrieval in the scattering network by gradient backpropagation, we introduce *scale-rate DAFx*, a class of audio transformations expressed in the domain of time–frequency scattering coefficients. One example of scale-rate DAFx is chirp rate inversion, which causes each sonic event to be locally reversed in time while leaving the arrow of time globally unchanged. Over the past two years, our work has led to the creation of four electroacoustic pieces: *FAVN*; *Modulator (Scattering Transform)*; *Experimental Palimpsest*; *Inspection (Maida Vale Project)* and *Inspection II*; as well as *XAllegroX (Hecker Scattering.m Sequence)*, a remix of Lorenzo Senni's *XAllegroX*, released by Warp Records on a vinyl entitled *The Shape of RemiXXXes to Come*.

1. INTRODUCTION

Several composers have pointed out the lack of a satisfying trade-off between interpretability and flexibility in the parametrization of sound transformations [1, 2, 3]. For example, the constant-Q wavelet transform (CQT) of an audio signal provides an intuitive display of its short-term energy distribution in time and frequency [4], but does not give explicit control over its intermittent perceptual features, such as roughness or vibrato. On the other hand, a deep convolutional generative model such as WaveNet [5] encompasses a rich diversity of timbre; but, because the mutual dependencies between the dimensions of its latent space are unspecified, music composition with autoencoders in the waveform domain is hampered by a long preliminary phase of trials and errors in the search for the intended effect.

Scattering transforms are a class of multivariable signal representations at the crossroads between wavelets and deep convolutional networks [6]. In this paper, we demonstrate that one such instance of scattering transform, namely time–frequency scattering [7], can be a relevant tool for composers of electroacoustic music, as it strikes a satisfying compromise between interpretability and flexibility. We describe the scattering-based DAFx underlying the synthesis of five electroacoustic pieces by Florian Hecker: *FAVN* (2016); *Modulator (Scattering Transform)* (2016-2017); *Experimental Palimpsest* (2016); *Inspection (Maida Vale Project)*(2016) and *Inspection II* (2017); as well as *XAllegroX (Hecker Scattering.m Sequence)*, a remix of Lorenzo Senni's *XAllegroX*, released by Warp Records on a vinyl entitled *The Shape of RemiXXXes to Come* (2017). In addition, we demonstrate the result of this algorithm in the companion website of this paper, which contains short audio examples as well as links to full-length computer-generated sound pieces.

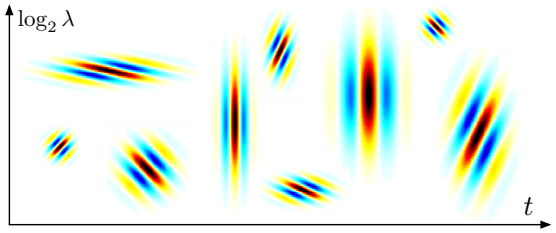


Figure 1: Interference pattern between wavelets $\psi_\alpha(t)$ and $\psi_\beta(\log_2 \lambda)$ in the time–frequency domain $(t, \log_2 \lambda)$ for different combinations of amplitude modulation rate α and frequency modulation scale β. Darker shades of red (resp. blue) indicate higher positive (resp. lower negative) values of the real part. See Section 2 for details.

Section 2 defines time–frequency scattering. Section 3 presents a gradient backpropagation method for sound synthesis from time–frequency scattering coefficients. Section 4 introduces "scale-rate DAFx", a new class of DAFx which operates in the domain of spectrotemporal modulations, and describes the implementation of chirp reversal as a proof of concept.

2. TIME–FREQUENCY SCATTERING

In this section, we define the time–frequency scattering transform as a function of four variables — time t, frequency λ, amplitude modulation rate α, and frequency modulation scale β — which we connect to spectrotemporal receptive fields (STRF) in auditory neurophysiology [8]. We refer to [9] for an in-depth mathematical introduction to time–frequency scattering.

This work is supported by the ERC InvariantClass grant 320959. The source code to reproduce experiments and figures is made freely available at: https://github.com/lostanlen/scattering.m

Companion website: https://lostanlen.com/pubs/dafx2019

2.1. Spectrotemporal receptive fields

Time–frequency scattering results from the cascade of two stages: a constant-Q wavelet transform (CQT) and the extraction of spectrotemporal modulations with wavelets in time and log-frequency. First, we define Morlet wavelets of center frequency $\lambda > 0$ and quality factor Q as

$$\boldsymbol{\psi}_\lambda(t) = \lambda \exp\left(-\frac{\lambda^2 t^2}{2Q^2}\right) \times (\exp(2\pi \mathrm{i} \lambda t) - \kappa), \qquad (1)$$

where the corrective term κ ensures that each $\boldsymbol{\psi}_\lambda(t)$ has one vanishing moment, i.e. a null average. In the sequel, we set $Q = 12$ to match twelve-tone equal temperament. Within a discrete setting, acoustic frequencies λ are typically of the form $2^{n/Q}$ where n is integer, thus covering the hearing range. For $\boldsymbol{x}(t)$ a finite-energy signal, we define the CQT of $\boldsymbol{x}$ as the matrix

$$\mathbf{U}_1\boldsymbol{x}(t, \lambda) = |\boldsymbol{x} * \boldsymbol{\psi}_\lambda|\,(t), \qquad (2)$$

that is, stacked convolutions with all wavelets $\boldsymbol{\psi}_\lambda(t)$ followed by the complex modulus nonlinearity.

Secondly, we define Morlet wavelets of respective center frequencies $\alpha > 0$ and $\beta \in \mathbb{R}$ with quality factor $Q = 1$. With a slight abuse of notation, we denote these wavelets by $\boldsymbol{\psi}_\alpha(t)$ and $\boldsymbol{\psi}_\beta(\log \lambda)$ even though they do not necessarily have the same shape as the wavelets $\boldsymbol{\psi}_\lambda(t)$ of Equation 2. Frequencies α, hereafter called amplitude modulation *rates*, are measured in Hertz (Hz) and discretized as 2^n with integer n. Frequencies β, hereafter called frequency modulation *scales*, are measured in cycles per octave (c/o) and discretized as $\pm 2^n$ with integer n. The edge case $\beta = 0$ corresponds to $\boldsymbol{\psi}_\beta(\log \lambda)$ being a Gaussian low-pass filter $\boldsymbol{\phi}_F(\log \lambda)$ of bandwidth F^{-1}. These modulation scales β play the same role as the *quefrencies* in a mel-frequency cepstrum [7].

We define the spectrotemporal receptive field (STRF) of $\boldsymbol{x}$ as the fourth-order tensor

$$\begin{aligned}\mathbf{U}_2\boldsymbol{x}(t, \lambda, \alpha, \beta) &= \left|\mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_\alpha \overset{\log_2 \lambda}{*} \boldsymbol{\psi}_\beta\right|(t, \lambda) \\ &= \left| \iint \mathbf{U}_1\boldsymbol{x}(\tau, s) \boldsymbol{\psi}_\alpha(t - \tau) \boldsymbol{\psi}_\beta(\log_2 \lambda - s)\, \mathrm{d}\tau\, \mathrm{d}s \right|, \qquad (3)\end{aligned}$$

that is, stacked convolutions in time and log-frequency with all wavelets $\boldsymbol{\psi}_\alpha(t)$ and $\boldsymbol{\psi}_\beta(\log_2 \lambda)$ followed by the complex modulus nonlinearity [10]. Figure 1 shows the interference pattern of the product $\boldsymbol{\psi}_\alpha(t - \tau)\boldsymbol{\psi}_\beta(\log_2 \lambda - s)$ for different combinations of time t, frequency λ, rate α, and scale β. We denote the multiindices (λ, α, β) resulting from such combinations as scattering *paths* [11]. We refer to [12] for an introduction to STRFs in the interdisciplinary context of music cognition and music information retrieval (MIR), and to [13] for an experimental benchmark in automatic speech recognition.

2.2. Invariance to translation

Because it is a convolutional operator in the time–frequency domain, the STRF is equivariant to temporal translation $t \mapsto t + \tau$ as well as frequency transposition $\lambda \mapsto 2^s\lambda$. In audio classification, it is useful to guarantee invariance to temporal translation up to some time lag T [11]. To this aim, we define time–frequency scattering as the result of a local averaging of both $\mathbf{U_1}\boldsymbol{x}(t, \lambda)$ and $\mathbf{U}_2\boldsymbol{x}(t, \lambda, \alpha, \beta)$ by a Gaussian low-pass filter $\boldsymbol{\phi}_T$ of cutoff frequency equal to T^{-1}, yielding

$$\mathbf{S}_1\boldsymbol{x}(t, \lambda) = \big(\mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\phi}_T\big)(t, \lambda) \quad \text{and} \qquad (4)$$

$$\mathbf{S}_2\boldsymbol{x}(t, \lambda, \alpha, \beta) = \big(\mathbf{U}_2\boldsymbol{x} \overset{t}{*} \boldsymbol{\phi}_T\big)(t, \lambda, \alpha, \beta) \qquad (5)$$

respectively. In practice, for purposes of signal classification, T is of the order of $50\,\mathrm{ms}$ in speech; of $500\,\mathrm{ms}$ in instrumental music; and of $5\,\mathrm{s}$ in ecoacoustics [14]. The delay of a real-time implementation of time–frequency scattering is of the order of T.

2.3. Energy conservation

We restrict the set of modulation rates α in $\mathbf{U_2}\boldsymbol{x}$ to values above T^{-1}, so that the power spectra of the low-pass filter $\boldsymbol{\phi}_T(t)$ and all wavelets $\boldsymbol{\psi}_\alpha(t)$ cover uniformly the Fourier domain [15, Chapter 4]: at every frequency ω, we have

$$\big|\widehat{\boldsymbol{\phi}}_T(\omega)\big|^2 + \frac{1}{2} \sum_{\alpha > T^{-1}} \Big(\big|\widehat{\boldsymbol{\psi}}_\alpha(\omega)\big|^2 + \big|\widehat{\boldsymbol{\psi}}_\alpha(-\omega)\big|^2\Big) \lessapprox 1, \qquad (6)$$

where the notation $A \lessapprox B$ indicates that there exists some $\varepsilon \ll B$ such that $B - \varepsilon < A < B$. In the Fourier domain associated to $\log_2 \lambda$, one has $\sum_\beta |\widehat{\boldsymbol{\psi}}_\beta(\omega)|^2 \lessapprox 1$ for all ω. Therefore, applying Parseval's theorem on all three wavelet filterbanks (respectively indexed by λ, α, and β) yields $\|\mathbf{S}_1\boldsymbol{x}\|_2^2 + \|\mathbf{U}_2\boldsymbol{x}\|_2^2 \lessapprox \|\mathbf{U}_1\boldsymbol{x}\|_2^2$. Figure 2 illustrates the design of these filterbanks in the Fourier domain.

The spectrotemporal modulations in music — e.g. tremolo, vibrato, and dissonance — are captured and demodulated by the second layer of a scattering network [16]. Consequently, each scattering path (λ, α, β) in $\mathbf{U}_2\boldsymbol{x}(t, \lambda, \alpha, \beta)$ yields a time series whose variations are slower than in the first layer $\mathbf{U}_1\boldsymbol{x}(t, \lambda)$; typically at rates of $1\,\mathrm{Hz}$ or lower. By setting T to 1 second or less, we may safely assume that the cutoff frequency of the low-pass filter $\boldsymbol{\phi}_T(t)$ in Equation 5 is high enough to retain all the energy in $\mathbf{U}_2\boldsymbol{x}(t, \lambda, \alpha, \beta)$. This assumption writes as $\|\mathbf{S}_2\boldsymbol{x}\| \lessapprox \|\mathbf{U}_2\boldsymbol{x}\|$ and is justified by the theorem of exponential decay of scattering coefficients [17]. Let $\mathbf{S}$ be the operator resulting from the concatenation of first-order scattering $\mathbf{S_1}$ with second-order scattering $\mathbf{S_2}$. In the absence of any DC bias in $\boldsymbol{x}(t)$, we conclude with the energy conservation identity

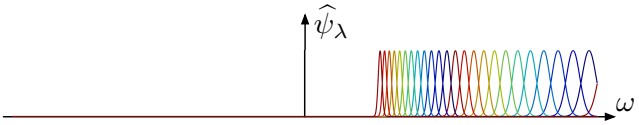
$$\|\mathbf{S}\boldsymbol{x}\|_2^2 = \|\mathbf{S}_1\boldsymbol{x}\|_2^2 + \|\mathbf{S}_2\boldsymbol{x}\|_2^2 \lessapprox \|\mathbf{U}_1\boldsymbol{x}\|_2^2 \lessapprox \|\boldsymbol{x}\|_2^2. \qquad (7)$$

3. AUDIO TEXTURE SYNTHESIS

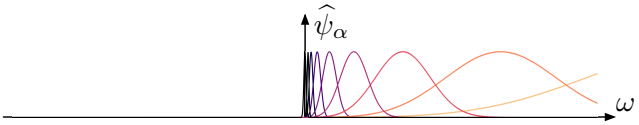
In this section, we describe how to pseudo-invert time–frequency scattering, that is, to generate a waveform whose scattering coefficients match the scattering coefficients of some other, pre-recorded waveform.

3.1. From phase retrieval to texture synthesis

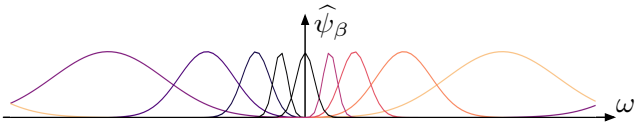
Although the invertibility of the convolutional operator involved in the constant-Q transform is guaranteed by wavelet frame theory [15, Chapter 5], the complex modulus nonlinearity in Equation 2 raises a fundamental question: is it always possible to recover $\boldsymbol{x}$, up to a constant and therefore imperceptible phase shift, from the magnitudes of its wavelet coefficients $\mathbf{U}_1\boldsymbol{x}(t, \lambda)$? This question has



(a) Wavelets $\widehat{\psi}_\lambda$ of frequencies λ and quality factor $Q = 12$.



(b) Wavelets $\widehat{\psi}_\alpha$ of rates α and quality factor $Q = 1$.



(c) Wavelets $\widehat{\psi}_\beta$ of scales β and quality factor $Q = 1$.

Figure 2: Filterbanks of Morlet wavelets in the Fourier domain: (a) for CQT; (b) for STRF, temporal dimension; (c) for STRF, log-frequential dimension. See Section 2 for details.

recently been answered in the affirmative, suggesting that a redundant CQT should be preferred over critically sampled short-term Fourier transforms (STFT) when attempting to sonify spectrograms in the absence of any prior information on the phase of the target waveform [18].

By recursion over layers in the composition of wavelet modulus operators, the invertibility of wavelet modulus implies the invertibility of scattering transforms of infinite depth [19], up to a constant time shift of at most T. However, the time–frequency scattering network presented here has a finite number of layers (i.e., most usually two layers) and is therefore not exactly invertible. Because of the theorem of exponential decay of scattering coefficients [17], the residual energy that is present in deeper layers can be neglected on the condition that T is small enough. In the rest of this section, we set T to $186\,\mathrm{ms}$, which corresponds to 8192 samples at a sample rate of $44.1\,\mathrm{kHz}$. Since the unit circle is not a convex subset of $\mathbf{R}^2$, the optimization problem $\boldsymbol{y}^*(t) = \arg\min_{\boldsymbol{y}} \mathbf{E}(\boldsymbol{y})$ where $\mathbf{E}(\boldsymbol{y}) = \frac{1}{2}\|\mathbf{S}\boldsymbol{x} - \mathbf{S}\boldsymbol{y}\|$ is nonconvex; therefore, its loss surface $\mathbf{E}$ may have local minimizers in addition to the global minimizers of the form $\boldsymbol{y}^*_\tau(t) = \boldsymbol{x}(t-\tau)$ with $|\tau| < T$. As a consequence, we formulate the problem of audio texture synthesis in loose terms of perceptual similarity. We refer to [20] for a literature review on texture synthesis, and to [21] for a discussion of quantitative evaluation procedures.

Starting from a colored Gaussian noise $\boldsymbol{y}_0(t)$ whose power spectral density matches $\mathbf{S}_1\boldsymbol{x}(t,\lambda)$, we refine it by additive updates of the form $\boldsymbol{y}_{n+1}(t) = \boldsymbol{y}_n(t) + \boldsymbol{u}_n(t)$, where the term $\boldsymbol{u}_n(t)$ is defined recursively as $\boldsymbol{u}_n(t) = m \times \boldsymbol{u}_n(t) + \mu_n \boldsymbol{\nabla}\mathbf{E}(\boldsymbol{y}_n)(t)$. In subsequent experiments, the momentum term is fixed at $m = 0.9$ while the learning rate is initialized at $\mu_0 = 0.1$ and modified at every step according to a "bold driver" heuristic [22].

3.2. Gradient backpropagation in a scattering network

Like deep neural networks, scattering networks consist of the composition of linear operators (wavelet transforms) and pointwise nonlinearities (complex modulus). Consequently, the gradient $\mathbf{E}(\boldsymbol{y}_n)$ can be obtained by composing the Hermitian adjoints of these operator in the reverse order as in the direct scattering transform — a method known as backpropagation [23].

First, we backpropagate the gradient of Euclidean loss for second-order scattering coefficients:

$$\boldsymbol{\nabla}\mathbf{U_2}\boldsymbol{y}(t,\lambda,\alpha,\beta) = \Big((\mathbf{S_2}\boldsymbol{x} - \mathbf{S_2}\boldsymbol{y}) \overset{t}{*} \boldsymbol{\phi}\Big)(t,\lambda,\alpha,\beta). \quad (8)$$

Secondly, we backpropagate the second layer onto the first:

$$\boldsymbol{\nabla}\mathbf{U_1}\boldsymbol{y}(t,\lambda) = \Big((\mathbf{S_1}\boldsymbol{x} - \mathbf{S_1}\boldsymbol{y}) \overset{t}{*} \boldsymbol{\phi}\Big)(t,\lambda,\alpha,\beta) + \sum_{\alpha,\beta} \Re\left(\left[\frac{(\mathbf{U_1}\boldsymbol{y} \overset{t}{*} \bar{\boldsymbol{\psi}}_\alpha) \overset{\log\lambda}{*} \bar{\boldsymbol{\psi}}_\beta)}{|\mathbf{U_1}\boldsymbol{y} \overset{t}{*} \bar{\boldsymbol{\psi}}_\alpha) \overset{\log\lambda}{*} \bar{\boldsymbol{\psi}}_\beta|} \times \boldsymbol{\nabla}\mathbf{U_2}\boldsymbol{y}\right] \overset{t}{*} \boldsymbol{\psi}_\alpha \overset{\log\lambda}{*} \boldsymbol{\psi}_\beta\right)(t,\lambda,\alpha,\beta), \quad (9)$$

where the symbol $\Re(z)$ denotes the real part of the complex number z. Lastly, we backpropagate the first layer into the waveform domain:

$$\boldsymbol{\nabla}\mathbf{E}(\boldsymbol{y})(t) = \sum_\lambda \Re\left(\left[\frac{\boldsymbol{y} \overset{t}{*} \boldsymbol{\psi}_\lambda}{|\boldsymbol{y} \overset{t}{*} \boldsymbol{\psi}_\lambda|} \times \boldsymbol{\nabla}\mathbf{U_1}\boldsymbol{y}\right] \overset{t}{*} \boldsymbol{\psi}_\lambda\right)(t) \quad (10)$$

Time–frequency scattering bears a strong resemblance with the set of spectrotemporal summary statistics developed by [24] to model the perception of auditory textures in the central auditory cortex. A qualitative benchmark has shown that time–frequency scattering is advantaged if $\boldsymbol{x}(t)$ contains asymmetric patterns (e.g. chirps), but that the two representations perform comparably otherwise [14]. Nevertheless, time–frequency scattering is considerably faster: the numerical optimizations of wavelet transforms and the recursive structure of backpropagation allows time–frequency scattering (both forward and backward) to be on par with real time on a personal computer, i.e. about 20 times faster than the other implementation. Therefore, an audio snippet of a few seconds can be fully re-synthesized in less than a minute, which makes it relatively convenient for composing sound serendipitously.

3.3. Creation: *FAVN* (2016) and other pluriphonic pieces

FAVN is an electroacoustic piece that evokes issues surrounding late 19th-century psychophysics as well as Debussy's *Prélude à l'après-midi d'un faune* (1894), which itself is a musical adaption of Stéphane Mallarmé's *L'après-midi d'un faune* (1876). To create *FAVN*, we began by composing 47 blocks of sound of duration equal to 21 seconds, and spatialized across three audio channels. These blocks are not directly created with time–frequency scattering; rather, they originate from the tools of the electroacoustic studio, such as oscillators and modulators. After digitizing these blocks, we analyze them and re-synthesize them by means of time–frequency scattering. We follow the algorithm described above: once initialized with an instance of Gaussian noise, the reconstruction is iteratively updated by gradient descent with a bold driver learning rate policy. We stop the algorithm after 50 iterations.

During the concert, the performer begins by playing the first iteration of the first block, and progressively moves forward in the reproduction of the piece, both in terms of compositional time (blocks) and computational time (iterations). The Alte Oper Frankfurt, Frankfurt am Main, Germany premiered *FAVN* on October

5th, 2016. The piece was presented again at *Geometry of Now* in Moscow in February 2017, and became a two-month exhibition at the Kunsthalle in Wien in November 2017, with a dedicated retrospective catalogue [25].

In the liner notes of *FAVN*, librettist Robin Mackay elucidates the crucial role of analysis-synthesis in the encoding of musical timbre:

> The analysis of timbre — a catch-all term referring to those aspects of the *thisness* of a sound that escape rudimentary parameters such a pitch and duration — is an active field of research today, with multiple methods proposed for classification and comparison. In *FAVN*, Hecker effectively reverses these analytical strategies devised for timbral description, using them to synthesize new sonic elements. In the first movement, a scattering transform with wavelets is used to produce an almost featureless ground from which an identifiable signal emerges as the texture is iteratively reprocessed to approximate its timbre. [Wavelets] correspond to nothing that can be heard in isolation, becoming perceptible only when assembled en masse. [26]

We refer to [27] for further discussions on the musical implications of time–frequency scattering, and to [28, 29] on the musical aesthetic of Florian Hecker. Since the creation of *FAVN*, we have used time–frequency scattering to create four new pieces.

Modulator (Scattering Transform) is a remix of Hecker's electronic piece *Modulator* (2012), obtained by retaining the 50^{th} iteration of the gradient descent algorithm. Editions Mego has released this remix in a stereo-cassette format. In addition, we presented an extended version of the piece in a 14-channel format at the exhibition "Florian Hecker - Formulations" at the Museum für Moderne Kunst Frankfurt, Frankfurt am Main, Germany from November 2016 to February 2017. In this multichannel version, 14 loudspeakers in the same room play a different iteration number of the reconstruction algorithm.

Experimental Palimpsest is an eight-channel variation upon *Palimpsest* (2004), Hecker's collaboration with the Japanese artist Yasunao Tone, obtained by the same procedure. This piece was premiered at the Lausanne Underground Film Festival, Lausanne, Switzerland, in October 2016.

Inspection (Maida Vale Project) is a seven-channel piece for synthetic voice and computer-generated sound, performed live at BBC's Maida Vale studios in London and has been broadcast on BBC Radio 3 in December 2016, marking the BBC's first ever live binaural broadcast. An extended version, *Inspection II*, will be released as a CD by Editions Mego, Vienna and Urbanomic, Falmouth, UK in Fall 2019.

3.4. *XAllegroX (scattering.m sequence)*

Lastly, *XAllegroX (scattering.m sequence)* is the remix of a dance music track by Lorenzo Senni, entitled *XAllegroX* and originally released by Warp Record in Senni's *The Shape of Trance to Come* LP (WAP406). Like in Hecker's experimental pieces, we remixed *XAllegroX* by, first, isolating a few one-bar loops from the original audio material, and secondly, by reconstructing them from their scattering coefficients. While, at the first iteration, the loop sounds hazy and static — or, a electronic musicians would call it, *droney* — it regains some of its original rhythmic content in subsequent iterations, thus producing a feeling of sonic rise that is fitting to the typical structuration of dance music. The peculiarity of this *scattering.m sequence* remix is that the musical "rise" is not produced over the well-known sonic attributes of frequency and amplitude (as is usually the case in electronic dance music), but on a relatively novel, joint parameter of texture; that is, a notion of sonic complexity which consists of the organization of frequencies and amplitude through time. Therefore, the development of gradient backpropagation for time–frequency scattering over new avenues for musical creation with digital audio effects: in addition to remixing amplitude (by EQ-ing) and frequency (by phase vocoder), it is now possible to *remix texture itself*, independently of amplitude and frequency. Along the same lines of amplitude modulation (AM) and frequency modulation (FM), we propose to call this new musical transformation a meta-modulation (MM), because it operates over spectrotemporal modulations rather than on the direct acoustic content. Future work will strive to further the understanding of MM, both from mathematical and compositional standpoints.

In July 2018, Warp released *XAllegroX (scattering.m sequence)* as part of a remix 12" named *The Shape of RemiXXXes to Come* (WAP425), hence the title of the present paper. This remix has been pressed on vinyl and made available on all major digital music platforms. The remix was aired on the Camarilha dos Quatro weekly podcast. Mary Anne Hobbs, an English DJ and music journalist, shared another of the album songs on her BBC show "6 Music Recommends". FACT listed the record as one of the must-haves of the month.

4. SCALE-RATE DIGITAL AUDIO EFFECTS

In this section, we introduce an algorithm to manipulate the finest time scales of spectrotemporal modulations (from $10\,\text{ms}$ to $1\,\text{s}$) while preserving both the temporal envelope and spectral envelope at a coarser scale (beyond $1\,\text{s}$). As an example, we implement chirp rate reversal, a new digital audio effect that flips the pitch contour of every note in a melody, without need for tracking partials. This concept will be featured throughout in the pluriphonic sound piece *Syn 21845* (2019), a sequel to Hecker's *Statistique Synthétique aux épaules de cascades* (2019).

4.1. Mid-level time scales in music perception

The invention of digital audio technologies allowed composers to apply so-called *intimate transformations* [30] to music signals, affecting certain time scales of sound perception while preserving others. The most prominent of such transformations is perhaps the phase vocoder [31], which transposes melodies and/or warps them in time independently. By setting T to $50\,\text{ms}$, a wavelet-based phase vocoder disentangles frequencies belonging to the hearing range (above $20\,\text{Hz}$) from modulation rates that are afferent to the perception of musical time (below $20\,\text{Hz}$) [32]. Frequency transposition is then formulated in $\mathbf{S}_1\boldsymbol{x}$ as a translation in $\log_2 \lambda$ whereas time stretching is formulated as a homothety in t.

In its simplest flavor, the phase vocoder suffers from artifacts near transient regions: because all time scales beyond T are warped in the same fashion, slowing down the tempo of a melody comes at the cost of a smeared temporal profile for each note. This well-known issue, which motivated the development of specific methods for transient detection and preservation [33], illustrates the importance of mid-level time scales in music perception, longer than a physical pseudo-period yet shorter than the time span between adjacent onsets [34].

The situation is different in a time–frequency scattering network: the amplitude modulations caused by sound transients are encoded in the scale-rate plane (α, β) of spectrotemporal receptive fields [7]. Therefore, time–frequency scattering appears as a convenient framework to address the preservation of such mid-level time scales in conjunction with a change in rhythmic parameters (meter and tempo); or, conversely, changes in articulation in conjunction with a preservation of the sequentiality in musical events.

4.2. General formulation

We propose to call *scale-rate DAFx* the class of audio transformations whose control parameters are foremostly expressed in the domain $(t, \lambda, \alpha, \beta)$ of time–frequency scattering coefficients, and subsequently backscattered to the time domain by solving an optimization problem of the form

$$\boldsymbol{y}^* = \arg\min_{y} \left\| f(\mathbf{S})\boldsymbol{x} - \mathbf{S}\boldsymbol{y} \right\|_2^2, \tag{11}$$

where the functional $f(\mathbf{S}) = (f_1(\mathbf{S}_1), f_2(\mathbf{S}_2))$ is defined by the composer. Compared to Section 3.1, the loss function in the equation above is not only nonconvex, but also devoid of a trivial global minimizer. Indeed, if the image of the reproducing kernel Hilbert space (RKHS) associated to Equation 3 by the complex modulus operator and low-pass filter $\boldsymbol{\phi}_T(t)$ (Equation 5) does not contain the function $f(\mathbf{S}\boldsymbol{x})$, then there is no constant-Q transform $\mathbf{U}_1^*(t, \log_2 \lambda)$ whose smoothed STRF is $f_2(\mathbf{S}_2)$; and *a fortiori* no waveform $\boldsymbol{y}^*(t)$ such that $\mathbf{S}\boldsymbol{y} = f(\mathbf{S})\boldsymbol{x}$. In order to allow for more flexibility in the set of valid choices of f, we replace the definition of $\mathbf{S}_2\boldsymbol{x}$ in Equation 5 by

$$\mathbf{S}_2\boldsymbol{x}(t, \lambda, \alpha, \beta) = \left(\mathbf{U}_2\boldsymbol{x} \overset{t}{*} \boldsymbol{\phi}_T \overset{\log_2 \lambda}{*} \boldsymbol{\phi}_F\right)(t, \lambda, \alpha, \beta), \tag{12}$$

that is, a blurring over both time and frequency dimensions; and likewise for $\mathbf{S}_1\boldsymbol{x}$. This new definition guarantees that $\mathbf{S}\boldsymbol{x}$ is invariant to frequency transposition up to intervals of size F (expressed in octaves), a property that is often desirable in audio classification [35]. Transposition-sensitive scattering (Equations 4 and 5) are a particular case of transposition-invariant scattering (equation above) at the $F \to 0$ limit, i.e. the Gaussian $\boldsymbol{\phi}_F$ becoming a Dirac delta distribution.

A thorough survey of scale-rate DAFx is beyond the scope of this article; in the sequel, we merely give some preliminary insights regarding their capabilities and limitations as well as a proof of concept. With $Q \gg 12$ wavelets per octave in the constant-Q transform and F of the order of one semitone, scale-rate DAFx would fall within the well-studied application domain of vibrato transformations [36]: a translation of the variable $\log_2 \alpha$ (resp. $\log_2 |\beta|$) would cause a multiplicative change in vibrato rate (resp. depth). Perhaps more interestingly, with $Q \ll 12$ and F of the order of an octave, scale-rate DAFx address the lesser-studied problem of roughness transformations in complex sounds: since the scattering transform captures pairwise interferences between pure tones within an interval of Q^{-1} octaves or less [16], a translation of the variable $(\log_2 \lambda + \log_2 \alpha)$ would transpose the sound while preserving its roughness, whereas a translation of the variable $(\log_2 \lambda - \log_2 \alpha)$ would affect roughness while preserving the spectral centroid. We believe that the capabilities of such transformations, both from computational and musical perspectives, are deserving of further inquiry.

4.3. Example: controlling the axis of time with chirp inversion

Because both Morlet wavelets $\boldsymbol{\psi}_\alpha(t)$ and $\boldsymbol{\psi}_\beta(\log_2 \lambda)$ have a symmetric profile, we have the following identity between Kronecker tensor products:

$$\boldsymbol{\psi}_\alpha \otimes \boldsymbol{\psi}_{-\beta} = \overline{\boldsymbol{\psi}_{-\alpha} \otimes \boldsymbol{\psi}_\beta}. \tag{13}$$

Since the constant-Q transform modulus $\mathbf{U}_1\boldsymbol{x}$ is real-valued, the above implies that $\mathbf{S}_2\boldsymbol{x}(t, \lambda, \alpha, -\beta) = \mathbf{S}_2\boldsymbol{x}(t, \lambda, -\alpha, \beta)$. In other words, flipping the sign of the modulation scale β is equivalent to reversing the time axis in the wavelet $\boldsymbol{\psi}_\alpha$; or, again equivalently, to reversing the time axis in the constant-Q transform $\mathbf{U}_1\boldsymbol{x}$ around the center of symmetry t before analyzing it with $\boldsymbol{\psi}_\alpha$ and $\boldsymbol{\psi}_\beta$. From these observations, we define a chirp inversion functional $f(\mathbf{S}) = (f_1(\mathbf{S}_1), f_2(\mathbf{S}_2))$ where $f_1(\mathbf{S}_1) = \mathbf{S_1}$ and f_2 is parameterized as

$$\begin{aligned} f_2 : \mathbf{S}_2(t, \lambda, \alpha, \beta) \longmapsto \quad & \frac{1 + \boldsymbol{\sigma}(t)}{2} \times \mathbf{S}_2(t, \lambda, \alpha, \beta) \\ & + \frac{1 - \boldsymbol{\sigma}(t)}{2} \times \mathbf{S}_2(t, \lambda, \alpha, -\beta), \end{aligned} \tag{14}$$

with $\boldsymbol{\sigma}(t)$ a slowly varying function at the typical time scale T. Observe that setting $\boldsymbol{\sigma}(t) = 1$ leaves $\mathbf{S}$ unchanged; that $\boldsymbol{\sigma}(t) = -1$ resembles short-time time reversal (STTR) of $\boldsymbol{x}(t)$ with half-overlapping windows of duration T [37]; and that $\boldsymbol{\sigma}(t) = 0$ produces a re-synthesized sound that is stationary, yet not necessarily Gaussian, up to the time scale T. It thus appears that the parameter $\boldsymbol{\sigma}(t)$ in Equation 14 is amenable to an "axis of time" knob that can be varied continuously through time within the range $[-1;\ 1]$.

As a proof of concept, Figure 3a shows the constant-Q transform of a repetitive sequence of synthetic chirps with varying amplitudes, frequential extents, and orientations; as well as its transformation by the functional f described above, with

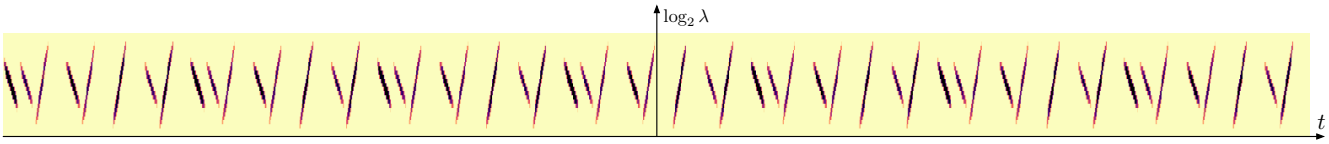
$$\boldsymbol{\sigma}(t) = \frac{1 - \exp\left(\frac{t}{\tau}\right)}{1 + \exp\left(\frac{t}{\tau}\right)} \tag{15}$$

the sigmoid function with a time constant $\tau \gg T$. The frequency transposition invariance F is set to 1 octave. We observe that, while the metrical structure of the original excerpt is recognizable at all times, the pitch contour of every musical event is identical to the original for $t \ll -\tau$ and inverted with respect to the original for $t \gg \tau$. For $|t| < \tau$, there is a progressive metamorphosis between the "forward time" and "backward time" regimes. The effect obtained in Figure 3b, although relatively simple to express in the scale-rate domain, would be difficult to implement in the time–frequency domain.

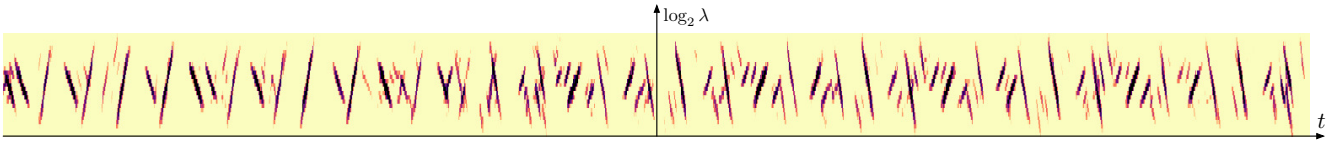
4.4. Towards digital audio effects on the pitch spiral

One evident drawback of scale-rate DAFx is the need to manually adjust the frequency transposition invariance F according to the analysis-synthesis task at hand. Forgoing this calibration step would certainly streamline the creative process. Furthermore, setting F to any value above 1 octave does not only affect spectrotemporal modulations but also the spectral envelope of $\mathbf{U}_1\boldsymbol{x}(t, \lambda)$. In the context of speech transformations, this undesirable phenomenon has led to the development of specific improvements to the phase vocoder [31].

With the aim of addressing both of these shortcomings, we suggest replacing the resort to STRF in Equation 3 $\mathbf{U}_2\boldsymbol{x}(t, \lambda)$ by



(a) An interpretation of Steve Reich's *Clapping Music* (1972) with synthetic chirps of varying rates, scales, and amplitudes.



(b) Re-synthesis after chirp rate inversion. The arrow of time goes forward for $t < 0$ and locally backward for $t > 0$.

Figure 3: An example of chirp rate inversion with time–frequency scattering. Top: original audio material. Bottom: computer-generated output after 100 iterations of gradient descent on time–frequency scattering coefficients. The chirp inversion functional follows a sigmoidal dynamic, as described in Equations 14 and 15. See Section 4 for details.

spiral scattering [38], a convolutional operator cascading wavelet transforms along time, along log-frequencies, and across neighboring octaves. Denoting by $\lfloor \log_2 \lambda \rfloor$ the octave index associated to the frequency λ – that is, the integer part of its binary logarithm – the spiral scattering transform of $\boldsymbol{x}(t)$ writes as

$$\mathbf{U}_2\boldsymbol{x}(t,\lambda,\alpha,\beta,\gamma) = \left|\mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_\alpha \overset{\log_2 \lambda}{*} \boldsymbol{\psi}_\beta \overset{\lfloor \log_2 \lambda \rfloor}{*} \boldsymbol{\psi}_\gamma\right|(t,\lambda) \quad (16)$$

where $\boldsymbol{\psi}_\gamma$ is a Morlet wavelet of quality factor $Q = 1$ and center frequency γ, with $|\gamma| < \frac{1}{2}$ measured in cycles per octave. Since spiral scattering disentangles temporal modulations of the non-stationary source-filter model [14], it is conceivable that source modulations and filter modulations could be manipulated independently in the space of spiral scattering coefficients. In particular, the aforementioned effect of "chirp rate reversal" could be generalized to the modulations of the source-filter model. For nonstationary harmonic tones, this would result in a reversal of melodic profile with preservation of the formantic profile, or vice versa. Although the present article does not give a demonstration of such effects, it is worth remarking that their future implementation in the *scattering.m* library would rely on the same principles as the gradient backpropagation of time–frequency scattering coefficients.

The procedure of rolling up the log-frequency axis into a spiral which makes a full turn at every octave, thus aligning power-of-two harmonics onto the same radius, is central to the construction of auditory paradoxes of pitch perception [39] and has recently been applied to musical instrument classification [14] and real-time pitch tuning [40]; yet, to the best of our knowledge, never as a mid-level representation for DAFx. As such, the theoretical framework between scale-rate DAFx and spiral DAFx lies at the interaction between two concurrent approaches in the DAFx community: sinusoidal modeling [33] and texture modeling via neural networks [41]. The former is more physically interpretable requires no training data, yet makes strong assumptions on the detectability of partials in the input spectrum; conversely, the latter is bereft of partial tracking, yet requires a training set and allows for less post hoc manipulations. The long-term goal of scale-rate DAFx is to borrow from both of these successful approaches, and ultimately achieve a satisfying compromise between interpretability and flexbility in texture synthesis.

5. CONCLUSION

The past decade has witnessed a breakthrough of deep convolutional architectures for signal classification, with some noteworthy applications in speech, music, and ecoacoustics. Yet there is, to this day, virtually no adoption of any recent deep learning system by electroacoustic music composers. This is due to several shortcomings of deep learning in its current state, among which: its high computational cost [42]; the need for a large dataset of musical samples, often supplemented with univocal human annotation [5]; the difficulty of synthesizing audio without artifacts [43]; and a certain opacity in the structure of the learned representation [44].

In this article, we have argued that time–frequency scattering – a deep convolutional operator involing little or no learning – is adequate for several use cases of contemporary music creation. We have supported our argumentation by three mathematical properties, which are rarely guaranteed in deep learning: energy conservation (Section 2); well-conditioned adjoint operators in closed form (Section 3); and psychophysiological interpretation in terms of modulation rates and scales (Section 4)..

All of the numerical applications presented here might, after enough effort, be implemented without resorting to time–frequency scattering at all. Yet, one noteworthy trait of time–frequency scattering resides in its versatility: it connects various topics of DAFx that are seemingly disparate, such as coding [4], texture synthesis [20], adaptive transformations [33], and similarity retrieval [45]. The guiding thread between these topics is that – adopting the categories of Iannis Xenakis [46] – time–frequency scattering extracts musical information at the *meso-scale* of musical motifs, allowing to put it in relation with both the *micro-scale* of musical timbre and the *macro-scale* of musical structures [28].

From a compositional perspective, the appeal of time–frequency scattering stems from the possibility to generate sound with a holistic approach, devoid of intermediate procedures for parametric estimation; yet leaving some room to serendipity and surprise in the listening experience. The best evidence of this fruitful trade-off between flexibility and interpretability is found in the breadth of computer music pieces that have resulted from the ongoing interaction between the two authors of this paper. The earliest musical creation that was based on time–frequency scattering (e.g. *FAVN* in 2016) was conceived as an operatic experience with pluriphonic spatialization at the Alte Oper Frankfurt, Frankfurt am Main,

Germany. In contrast, *Modulator (Scattering Transform)* (2017) is a stereo-cassette mix for Editions Mego; *Inspection (Maida Vale Project)* (2017) is a live radio performance at the BBC; and lastly, *XAllegroX (Hecker Scattering.m Sequence)* (2018) is the remix of a dance music track for Warp Records.

More than a fortuitous affinity of personal initiatives, the research-creation agenda that is outlined in the present paper wishes to espouse the noble tradition of *musical research* [47], i.e. a kind of creative process in which the conventional division of labor between scientists and artists is tentatively called into question. Here, for the sake of crafting new music, a signal processing researcher (VL) becomes *de facto* a computer music designer, while a composer (FH) takes on a role akin to a principal investigator [48].

6. ACKNOWLEDGMENTS

This work is supported by the ERC InvariantClass grant 320959 and NSF awards 1633259 and 1633206. The two authors wish to thank Bob Sturm for putting them in contact with each other; Lorenzo Senni for accepting that his record title, *The Shape of RemiXXXes to Come*, is being reused as the title of the present article; and the anonymous reviewers of both DAFX-18 and DAFX-19 for their helpful comments.

7. REFERENCES

[1] H Kaper and S Tipei, "Formalizing the concept of sound," in *Proc. ICMC*, 1999.

[2] J.-C Risset, "Fifty years of digital sound for music," in *Proc. SMC*, 2007.

[3] C.-E Cella, "Machine listening intelligence," in *Proc. Int. Workshop on Deep Learning for Music*, 2017.

[4] G. A Velasco, N Holighaus, M Dörfler, and T Grill, "Constructing an invertible constant-Q transform with non-stationary Gabor frames," in *Proc. DAFx*, 2011.

[5] J Engel, C Resnick, A Roberts, S Dieleman, D Eck, K Simonyan, and M Norouzi, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proc. ICML*, 2017.

[6] S Mallat, "Understanding deep convolutional networks," *Phil. Trans. R. Soc. A*, vol. 374, no. 2065, 2016.

[7] J Andén, V Lostanlen, and S Mallat, "Joint time-frequency scattering for audio classification," in *Proc. MLSP*. IEEE, 2015, pp. 1–6.

[8] K Patil, D Pressnitzer, S Shamma, and M Elhilali, "Music in our ears: the biological bases of musical timbre perception," *PLoS computational biology*, vol. 8, no. 11, 2012.

[9] J Andén, V Lostanlen, and S Mallat, "Ieee trans. sig. proc.," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3704–3718, July 2019.

[10] T Lindeberg and A Friberg, "Idealized computational models for auditory receptive fields," *PLoS one*, vol. 10, no. 3, 2015.

[11] S Mallat, "Group invariant scattering," *Comm. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.

[12] K Siedenburg, I Fujinaga, and S McAdams, "A comparison of approaches to timbre descriptors in music information retrieval and music psychology," *J. New Music Research*, vol. 45, no. 1, pp. 27–41, 2016.

[13] M. R Schädler, B. T Meyer, and B Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. of Am.*, vol. 131, no. 5, pp. 4134–4151, 2012.

[14] V Lostanlen, *Convolutional operators in the time-frequency domain*, Ph.D. thesis, École normale supérieure, 2017.

[15] S Mallat, *A wavelet tour of signal processing: the sparse way*, Academic press, 2008.

[16] J Andén and S Mallat, "Scattering representation of modulated sounds," in *Proc. DAFx*, 2012.

[17] I Waldspurger, "Exponential decay of scattering coefficients," in *Proc. IEEE SampTA*, 2017.

[18] I Waldspurger, "Phase retrieval for wavelet transforms," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 2993–3009, 2017.

[19] I Waldspurger, *Wavelet transform modulus: phase retrieval and scattering*, Ph.D. thesis, École normale supérieure, 2015.

[20] D Schwarz, "State of the art in sound texture synthesis," in *Proc. DAFx*, 2011.

[21] D Schwarz, A Röbel, C Yeh, and A Laburthe, "Concatenative sound texture synthesis methods and evaluation," in *Proc. DAFx*, 2016.

[22] I Sutskever, J Martens, G Dahl, and G Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. ICML*, 2013, pp. 1139–1147.

[23] J Bruna and S Mallat, "Audio texture synthesis with scattering moments," *arXiv preprint arXiv:1311.0407*, 2013.

[24] J. H McDermott and E. P Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[25] V. J. M Nicolaus Schafhausen, Ed., *Florian Hecker: Halluzination, Perspektive, Synthese*, Sternberg Press, Berlin, 2019.

[26] R Mackay, Program notes to *FAVN*'s premiere. Alte Oper, Frankfurt, October 5th, 2016.

[27] V Lostanlen, "On time-frequency scattering and computer music," in *Florian Hecker: Halluzination, Perspektive, Synthese*, N Schafhausen and V. J Müller, Eds. Sternberg Press, Berlin, 2019.

[28] F Hecker and R Mackay, "Sound out of line: In conversation with Florian Hecker," *Urbanomic*, , no. 3, 2009.

[29] F Hecker, Ed., *Chimerizations*, Primary Information, New York, 2013.

[30] J.-C Risset, "Exploration of timbre by analysis and synthesis," in *The Psychology of Music, 2nd Ed.*, D Deutsch, Ed., chapter 5, pp. 113–169. Elsevier, 1999.

[31] A Röbel, "A shape-invariant phase vocoder for speech transformation," in *Proc. DAFx*, 2010.

[32] R Kronland-Martinet, "The wavelet transform for analysis, synthesis, and processing of speech and music sounds," *Comp. Mus. J.*, vol. 12, no. 4, pp. 11–20, 1988.

[33] A Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. DAFX*, 2003.

[34] P Leveau, E Vincent, G Richard, and L Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 16, no. 1, pp. 116–128, 2008.

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

[36] A Röbel, S Maller, and J Contreras, "Transforming vibrato extent in monophonic sounds," in *Proc. DAFx 2011*, 2011.

[37] H.-S Kim and J. O. I Smith, "Short-time time-reversal on audio signals," in *Proc. DAFx*, 2014.

[38] V Lostanlen and S Mallat, "Wavelet scattering on the pitch spiral," in *Proc. DAFx*, 2015.

[39] 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.

[40] T Hélie and C Picasso, "The Snail: a real-time software application to visualize sounds," in *Proc. DAFx*, 2017.

[41] H Caracalla, "Synthèse de textures sonores à partir de statistiques temps-fréquence," M.S. thesis, Ircam, 2016.

[42] M Kim and P Smaragdis, "Bitwise neural networks," in *Proc. ICML*, 2015.

[43] C Donahue, J Macaulay, and M Puckette, "Synthesizing audio with GANs," in *Proc. ICLR, workshop track*, 2018.

[44] G Lample, N Zeghidour, N Usunier, A Bordes, L Denoyer, et al., "Fader networks: Manipulating images by sliding attributes," in *Proc. NIPS*, 2017.

[45] T Pohle, E Pampalk, and G Widmer, "Generating similarity-based playlists using traveling salesman algorithms," in *Proc. DAFx*, 2005.

[46] I Xenakis, "Concerning time," *Perspectives of New Music*, vol. 1, no. 27, pp. 84–92, 1989.

[47] J.-C Risset, "Le compositeur et ses machines : de la recherche musicale," *Esprit*, vol. 3, no. 99, pp. 59–76, 1985.

[48] A Cont and A Gerzso, "The C in IRCAM: Coordinating musical research at IRCAM," in *Proc. ICMC*, 2010.

[49] R Mackay, Ed., *Florian Hecker: Formulations*, Koenig Books, London, 2016.

[50] V Lostanlen, "Scattering.m: a matlab toolbox for wavelet scattering," June 2019.