# ELECTROACOUSTIC MUSIC CREATIONS WITH TIME-FREQUENCY SCATTERING

*Vincent Lostanlen*
Music and Audio Research Lab
New York University
New York, NY, USA vincent.lostanlen@nyu.edu

*Florian Hecker*
Edinburgh College of Art
University of Edinburgh
Edinburgh, UK
florian.hecker@ed.ac.uk

## ABSTRACT

VL: Missing abstract

## 1. INTRODUCTION

Visualization and control are two of the main challenges facing the adoption of digital audio effects (DAFx) in contemporary music creation [1]. During the past two decades, extensive research has addressed each of these challenges independently; yet, rarely ever in combination. Several composers have pointed out the lack of a satisfying trade-off between interpretability and flexibility in the parametrization of sound transformations [2, 3, 4]. 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 [5], 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 [6] 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 [7]. In this paper, we demonstrate that one such instance of scattering transform, namely time-frequency scattering [8], 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 pieces by Florian Hecker: *FAVN* (2016), *Modulator (Scattering Transform)* (2016-2017), *Experimental Palimpsest* (2016), *Inspection* (2016-2017), and *Synopsis Seriation* (2018). In particular, we show that the resort to time-frequency scattering goes well beyond the manipulation of timbre ; it also gives an insight on other aspects of composition, including motivic variation, anticipation, spatialized pluriphony, and the relationships between sound and text.

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. Lastly, Section 5 describes an algorithm for the segmentation and re-ordering of time-frequency scattering representations based on information geometry (generalized likelihood ratios) and combinatorial optimization (travelling salesman problem). Each section concludes with a survey of the music creations derived from it.

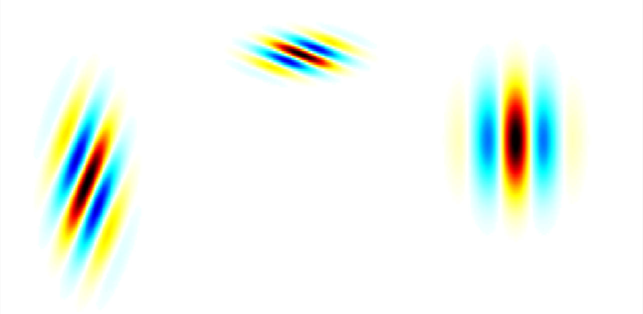

Figure 1: Interference pattern between wavelets $\boldsymbol{\psi}_\alpha(t)$ and $\boldsymbol{\psi}_\beta(\log_2 \lambda)$ in the time-frequency domain $(t, \log_2 \lambda)$ for different combinations of amplitude modulation rate $\alpha$ and frequency modulation scale $\beta$. Darker shades of red (resp. blue) indicate higher positive (resp. lower negative) values of the real part.VL: Axes

## 2. TIME-FREQUENCY SCATTERING

In this section, we define the time-frequency scattering transform as a function of four variables – time $t$, frequency $\lambda$, amplitude modulation rate $\alpha$, and frequency modulation scale $\beta$ – which we connect to spectrotemporal receptive fields (STRF) in auditory neurophysiology [9]. We describe how the property of energy conservation led to the publication of *Scattering to Text* (2018). VL: citation needed.

### 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 $\kappa$ 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 $\lambda$ 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 are not homothetic to the wavelets $\boldsymbol{\psi}_\lambda(t)$ of Equation 2. Frequencies $\alpha$, hereafter called amplitude modulation *rates*, are measured in Hertz (Hz) and discretized as $2^n$ with integer $n$. Frequencies $\beta$, sometimes called "quefrencies" and hereafter called frequency modulation *scales*, are measured in cycles per octave (c/o) and discretized as $\pm 2^n$ with integer $n$. In addition, the edge case $\beta = 0$ corresponds to $\boldsymbol{\psi}_\beta(\log \lambda)$ being a Gaussian lowpass filter $\boldsymbol{\phi}_F(\log \lambda)$ of bandwidth $F^{-1}$.

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

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

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 $\lambda$, rate $\alpha$, and scale $\beta$. We refer to [11] for an introduction to STRFs in the interdisciplinary context of music cognition and music information retrieval (MIR), and to [12] 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 $T$ [13]. 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 lowpass filter $\boldsymbol{\phi}_T$ of cutoff frequency equal to $T^{-1}$, yielding

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

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

respectively. In practice, $T$ is of the order of 50 ms in speech; of 500 ms in instrumental music; and of 5 s in ecoacoustics [14].

### 2.3. Energy conservation

We restrict the set of modulation rates $\alpha$ in $\mathbf{U}_2\boldsymbol{x}$ to values above $T^{-1}$, so that the power spectra of the lowpass filter $\boldsymbol{\phi}_T(t)$ and all wavelets $\boldsymbol{\psi}_\alpha(t)$ cover uniformly the Fourier domain [15, Chapter 4]: at every frequency $\omega$, we have

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

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 $\omega$. Therefore, applying Parseval's theorem on all three wavelet filterbanks (respectively indexed by $\lambda$, $\alpha$, and $\beta$) yields $\|\mathbf{S}_1\boldsymbol{x}\|_2^2 + \|\mathbf{U}_2\boldsymbol{x}\|_2^2 \lessapprox \|\mathbf{U}_1\boldsymbol{x}\|_2^2$.

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 $(\lambda,\alpha,\beta)$ in $\mathbf{U}_2\boldsymbol{x}(t,\lambda,\alpha,\beta)$ is a time series whose variations are slower than in the first layer $\mathbf{U}_1\boldsymbol{x}(t,\lambda)$; typically at rates of 1 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]. In the absence of any DC bias in $\boldsymbol{x}(t)$, we conclude with the energy conservation identity

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

### 2.4. Creation: *Scattering to Text* (2018)

## 3. AUDIO TEXTURE SYNTHESIS

### 3.1. From phase retrieval to texture synthesis

Until today, time-frequency scattering has exclusively been applied to audio classification tasks.

We refer to [18] for a literature review on texture synthesis, and to [19]

[20]

### 3.2. Gradient backpropagation in a scattering network

$$\boldsymbol{y}^*(t) = \arg\min_{y} \big\|\mathbf{S}\boldsymbol{x} - \mathbf{S}\boldsymbol{y}\big\| \quad (8)$$

$$\boldsymbol{\nabla}\mathbf{U}_2\boldsymbol{x}(t,\lambda,\alpha,\beta) = (\boldsymbol{\nabla}\mathbf{S}_2\boldsymbol{x} \overset{t}{*} \boldsymbol{\phi})(t,\lambda,\alpha,\beta) \quad (9)$$

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

A preliminary version of the piece is composed by means of computer music tools and mixed down to three monophonic channels, each segmented into 47 blocks of duration equal to 21 seconds. Each block of each channel is reconstructed independently with texture synthesis

We reproduce here an excerpt of the program notes, written by philosopher Robin MacKay:

> *FAVN* also folds Mallarmé's insistence on the impossibility of cataloguing the idea in plain prose onto Hecker's concern with the ways in which sound is analytically coded, in particular focusing on the concept of timbre [...]. The analysis of timbre – a catch-all term referring to those aspects of the *thisness* of a sound that escape rudimentary parameters such as 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 employed to produce an almost featureless ground from which an identifiable signal emerges as the texture is iteratively reprocessed to approximate its timbre. Rather than operating via the superposition of pure tones, wavelets furnish a kind of timbral dictionary; in themselves

> they correspond to nothing that can be heard in isolation, becoming perceptible only when assembled en masse – at which point one hears not distinct wavelets, but an emergent overall timbre. [21]

## 4. SCALE-RATE DIGITAL AUDIO EFFECTS

In this section, we introduce an algorithm to manipulate the finest time scales of spectrotemporal modulations (from $10\,\mathrm{ms}$ to $1\,\mathrm{s}$) while preserving both the temporal envelope and spectral envelope at a coarser scale (beyond $1\,\mathrm{s}$). 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.

### 4.1. Mid-level time scales in music perception

The invention of digital audio technologies allowed composers to apply so-called *intimate transformations* [22] to music signals, affecting certain time scales of sound perception while preserving others. The most prominent of such transformations is perhaps the phase vocoder [23], which transposes melodies and/or warps them in time independently. By setting $T$ to $50\,\mathrm{ms}$, a wavelet-based phase vocoder disentangles frequencies belonging to the hearing range (above $20\,\mathrm{Hz}$) from modulation rates that are afferent to the perception of musical time (below $20\,\mathrm{Hz}$) [24]. 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 [25], 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 [26].

The situation is different in a time-frequency scattering network: the amplitude modulations caused by sound transients are encoded in the scale-rate plane $(\alpha, \beta)$ of spectrotemporal receptive fields [8]. 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

$$\boldsymbol{y}^* = \arg\min_{y} \big\| f(\mathbf{S})\boldsymbol{x} - \mathbf{S}\boldsymbol{y} \big\|_2^2, \tag{10}$$

where the functional $f(\mathbf{S}) = (f_1(\mathbf{S}_1), f_2(\mathbf{S}_2))$ is defined by the composer. Compared to Equation 8, 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 lowpass 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

$$\mathbf{S}_2\boldsymbol{x}(t, \lambda, \alpha, \beta) = \big(\mathbf{U}_2\boldsymbol{x} \overset{t}{*} \boldsymbol{\phi}_T \overset{\log_2 \lambda}{*} \boldsymbol{\phi}_F\big)(t, \lambda, \alpha, \beta), \tag{11}$$

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 [27]. Transposition-sensitive scattering (Equations 4 and 5) are a particular case of transposition-invariant scattering (equation above) at the $F \rightarrow 0$ 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, they would fall within the well-studied application domain of vibrato transformations [28]: 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.

### 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:

$$\boldsymbol{\psi}_\alpha \otimes \boldsymbol{\psi}_{-\beta} = \overline{\boldsymbol{\psi}_{-\alpha} \otimes \boldsymbol{\psi}_\beta}. \tag{12}$$

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 $\beta$ 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

$$\begin{aligned} f_2 : \mathbf{S}_2(t, \lambda, \alpha, \beta) \longmapsto \quad & \frac{1 + \boldsymbol{\sigma}(t)}{2} \times \mathbf{S}_2(t, \lambda, \alpha, \beta) \\ & + \frac{1 - \boldsymbol{\sigma}(t)}{2} \times \mathbf{S}_2(t, \lambda, \alpha, -\beta), \end{aligned} \tag{13}$$

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 $T$ [29]; 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 13 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 VL: X 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

$$\boldsymbol{\sigma}(t) = \frac{1 - \exp\left(\frac{t}{\tau}\right)}{1 + \exp\left(\frac{t}{\tau}\right)} \tag{14}$$

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 VL: X, although relatively simple to express in the scale-rate domain, would be difficult to implement in the spectrogram domain.

### 4.4. Towards digital audio effects on the pitch spiral

One evident drawback of scale-rate DAFX is the need to manually tune the frequency transposition invariance $F$ according to the analysis-synthesis task at hand.

[23] [30]

$$\mathbf{U}_2\boldsymbol{x}(t, \lambda, \alpha, \beta, \gamma) = \left| \mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_\alpha \overset{\log_2 \lambda}{*} \boldsymbol{\psi}_\beta(t, \lambda) \overset{\lfloor \log_2 \lambda \rfloor}{*} \boldsymbol{\psi}_\gamma \right| (t, \lambda) \tag{15}$$

## 5. SEGMENTATION AND SERIATION

### 5.1. Real-time audio segmentation with generalized likelihood ratios

[31] [32] Both the parameters before and after change are unknown.

### 5.2. Online segmentation with generalized likelihood ratios

$$\mathbf{X}_{\mathrm{P}}(t_0, t_1) : (t, k) \longmapsto \sum_{\tau = t_0}^{t} \mathbf{S}\boldsymbol{x}[t, k]^2 \tag{16}$$

$$\mathbf{X}_{\mathrm{F}}(t_0, t_1) : (t, k) \longmapsto \sum_{\tau = t+1}^{t_1} \mathbf{S}\boldsymbol{x}[t, k]^2 \tag{17}$$

$$\mathbf{X}_{\mathrm{P}\cup\mathrm{F}}(t_0, t_1) = \mathbf{X}_{\mathrm{P}}(t_0, t_1) + \mathbf{X}_{\mathrm{F}}(t_0, t_1) \tag{18}$$

$$\begin{aligned} \log_2 \mathrm{GLM}(\boldsymbol{x}, t_0, t_1) = \; & (t_1 - t_0) \times H(\mathbf{X}_{\mathrm{P}\cup\mathrm{F}}(t_0, t_1))[t] \\ & - (t - t_0) \times H(\mathbf{X}_{\mathrm{P}}(t_0, t_1)) \\ & - (t_1 - t) \times H(\mathbf{X}_{\mathrm{F}}(t_0, t_1)), \end{aligned} \tag{19}$$

where $H$ is the Shannon entropy

$$H(\mathbf{X}) : t \longmapsto -\sum_{k=1}^{K} \mathbf{X}[t, k] \log_2 \mathbf{X}[t, k]. \tag{20}$$

$$\mathbf{L}[n_0, n_1] = -\mathrm{JS}(\mathbf{X}_{\mathbf{n_0}}, \mathbf{X}_{\mathbf{n_1}}) + \sum_{n=1}^{N} \mathrm{JS}(\mathbf{X}_{\mathbf{n_0}}, \mathbf{X}_{\mathbf{n}}) \tag{21}$$

$$\boldsymbol{v}^* = \arg \min_{\boldsymbol{v} \in \mathbb{R}^N} \left\{ \langle \boldsymbol{v} | \mathbf{L}\boldsymbol{v} \rangle \,\middle|\, \|\boldsymbol{v}\| = 1, \sum_{n=1}^{N} \boldsymbol{v}_n = 0 \right\} \tag{22}$$

$\boldsymbol{v}_{\pi(1)} < \ldots < \boldsymbol{v}_{\pi(N)}$
[33]

### 5.3. Creation: *Synopsis Seriation* (2018)

## 6. CONCLUSIONS

All of the applications presented here might, after enough effort, be implemented without resorting to time-frequency scattering. What is novel is how this framework connects various subfields of DAFx.

This paper is the result of an ongoing collaboration between a signal processing academic and a composer, fostering both research and creation.

## 7. ACKNOWLEDGMENTS

This work is supported by the ERC InvariantClass grant 320959, NSF awards 1633259 and 1633206, and a Google faculty award. The two authors wish to thank Bob Sturm for putting them in contact with each other.

## 8. REFERENCES

[1] Daniel Arfib, "Visual representations for digital audio effects and their control," in *Proc. DAFx*, 1999.

[2] Hans Kaper and Sever Tipei, "Formalizing the concept of sound," in *Proc. ICMC*, 1999.

[3] Jean-Claude Risset, "Fifty years of digital sound for music," in *Proc. SMC*, 2007.

[4] Carmine-Emanuele Cella, "Machine listening intelligence," in *Proc. Int. Workshop on Deep Learning for Music*, 2017.

[5] Gino Angelo Velasco, Nicki Holighaus, Monika Dörfler, and Thomas Grill, "Constructing an invertible constant-Q transform with non-stationary Gabor frames," in *Proc. DAFx*, 2011.

[6] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proc. ICML*, 2017.

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

[8] Joakim Andén, Vincent Lostanlen, and Stéphane Mallat, "Joint time-frequency scattering for audio classification," in *Proc. IEEE MLSP*, 2015.

[9] Kailash Patil, Daniel Pressnitzer, Shihab Shamma, and Mounya Elhilali, "Music in our ears: the biological bases of musical timbre perception," *PLoS computational biology*, vol. 8, no. 11, 2012.

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

[11] Kai Siedenburg, Ichiro Fujinaga, and Stephen McAdams, "A comparison of approaches to timbre descriptors in music information retrieval and music psychology," *Journal of New Music Research*, vol. 45, no. 1, pp. 27–41, 2016.

[12] Marc René Schädler, Bernd T. Meyer, and Birger 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.

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

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

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

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

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

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

[19] Diemo Schwarz, Axel Roebel, Chunghsin Yeh, and Amaury Laburthe, "Concatenative sound texture synthesis methods and evaluation," in *Proc. DAFx*, 2016.

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

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

[22] Jean-Claude Risset, "Exploration of timbre by analysis and synthesis," in *The Psychology of Music, 2nd Ed.*, Diana Deutsch, Ed., chapter 5, pp. 113–169. Elsevier, 1999.

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

[24] Richard 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.

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

[26] Pierre Leveau, Emmanuel Vincent, Gaël Richard, and Laurent Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 16, no. 1, pp. 116–128, 2008.

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

[28] Axel Roebel, Simon Maller, and Javier Contreras, "Transforming vibrato extent in monophonic sounds," in *Proc. DAFx 2011*, 2011.

[29] Hyung-Suk Kim and Julius Orion III Smith, "Short-time time-reversal on audio signals," in *Proc. DAFx*, 2014.

[30] Vincent Lostanlen and Stéphane Mallat, "Wavelet scattering on the pitch spiral," in *Proc. DAFx*, 2015.

[31] Arnaud Dessein and Arshia Cont, "An information-geometric approach to real-time audio segmentation," *IEEE Sig. Proc. Lett.*, vol. 20, no. 4, pp. 331–334, 2013.

[32] Vincent Lostanlen, "Découverte automatique de structures musicales en temps réel par la géometrie de l'information," M.S. thesis, Ircam, 2013.

[33] Geoffroy Peeters and Xavier Rodet, "Signal-based music structure discovery for audio summary generation," in *Proc. ICMC*, 2003.