The quest for an adequate representation of auditory textures lies at the foundation of computer music research. Indeed, none of its analog predecessors ever managed a practical compromise between two concurrent needs in sound design: first, to reproduce faithfully any pre-existing texture; and second, to offer enough flexibility for sculpting novel textures from scratch. For example, Schaeffer’s *musique concrète* offered a precise typology of musical objects, yet constrains the composer to a raw, figurativistic material. On the other hand, Stockhausen’s *Elektronische Musik*, as it arranges simple noises and tones through time, may have uncovered new avenues in musical abstraction; yet at the cost of a narrow, distinctively “robotic” timbral palette. In the history of music technology, such an opposition between specificity and expressivity is reflected in the respective developments of granular synthesis and additive synthesis: one is universal but computationally intractable; the other is terse but somewhat clunky. With the democratization of analog-to-digital audio conversion, both aforementioned schools of thought came into decline, and new tools for sound manipulation in the time-frequency domain, such as the phase vocoder, gained momentum among contemporary music composers. However, the progressive digitization of the music studio has brought little progress to the long-lasting problem of audio texture synthesis and manipulation.

The science of auditory neurophysiology paved the way towards a computational framework for audio texture modeling that could reconcile the specificity of *musique concrète* with the expressivity of *Elektronische Musik*. In 1996, Nina Kowalski and her colleagues employed an array of silicon electrodes to measure the cortical responses of a ferret to computer-generated ripple stimuli, exhibiting modulations in both time and frequency. Pairwise correlations between stimuli and responses led to an exhaustive mapping of the primary auditory cortex of mammals, which associates each neuron to a spectrotemporal receptive field (STRF)—that is, the time-frequency representation pattern eliciting maximal excitation of this neuron. Kowalski et al. concluded that our brain integrates the acoustic spectrum through time in terms of its spectrotemporal modulations at various scales (pitch intervals) and rates (pulse tempi). Neither exclusively rhythmic (temporal), nor exclusively harmonic (frequent), our brain is indeed a joint, rhythmico-harmonico-melodic processor that encodes sound into a multifaceted sensation.

Despite marking a watershed in our understanding of music perception, this finding long remained outside the technological landscape of computer music designers because the biologically inspired STRF representation was not an invertible procedure. Instead, although STRF allowed mapping sounds to specific areas of the auditory cortex, the dual problem of sonifying the neuro-electrical activations of these areas had remained largely unexplored. In addition, since STRF had been obtained empirically from ferret neuronal action potentials, the resulting representation could not be interpreted post hoc in terms of continuous perceptual parameters, such as pitch or tempo. Simply put, STRF are more concrete than *musique concrète* itself—in lieu of ear-drum vibrations, what they contain is a heatmap of primary auditory cortex activity—but lack the mathematical concision of an *Elektronische Musik* score in order to allow for any compositional intervention on the world of natural sounds.

From 2013 to 2016, I was a graduate student at École normale supérieure, striving to develop new convolutional operators in the time-frequency domain.  

---

domain for modeling musical timbre. With my coworker Joakim Andén and my advisor Stéphane Mallat, I contributed to a STRF-based computational model for audio texture synthesis, under the name of time-frequency scattering. Time-frequency scattering was formulated as the successor to “time scattering,” as it was formulated by Mallat himself in 2012. The name was coined as a nod to the world of quantum mechanics: from the reddish shade of a sunset to the glistening of a pearl, many are the microscopic phenomena encompassed by the umbrella term of scattering. The commonality between such phenomena is that they all involve a radiation of some kind as well as a maze of nonuniformities. Let $g$ be a Gaussian bell curve. In the context of scattering transforms, the radiation a sound pressure wave $U_{\alpha}(t)$, while the maze consists of Morlet wavelets

$$\psi_{\alpha}(t) = \lambda g(\lambda^2 t) \exp(2\pi i \lambda^2 t) - g(\lambda^2 t)$$

[1]

tuned at resolutions $\lambda$, as well as modulus nonlinearities.

Before time-frequency scattering was formalized, Mallat had defined the time scattering transform as a cascade of purely temporal wavelet modulus operators:

$$U_{\alpha}(t) = \int_{-\infty}^{\infty} U_{\alpha}(\tau, t) \psi_{\alpha}(\tau - t) \, d\tau$$

[2]

and then generalized his theory to all real-valued functions of finite energy defined over the irreducible representations of a given compact Lie group. Shortly thereafter, my coworker Irène Waldspurger proved that scattering transforms—despite the loss of the phase incurred by the complex moduli—are invertible with continuous inverse. She resorted to advanced methods in tomography and meromorphic extensions, among others, to come up with this astonishing result: on the condition that the wavelet transforms form a “tight” frame of the functional space at hand, and towards the limit of infinite depth $n \to \infty$, the time variable can ultimately be removed from the equation, because the oscillatory nature of sound vibrations in $U_{\alpha}(t)$ becomes fully characterized by its interference pattern through the scattering network. Going back to the metaphor of the Mie scattering in quantum mechanics, it is as though Mallat and Waldspurger had unearthed some kind of all-witnessing crystal, whose eternal glint was a petrified testimony of every light it had seen before.

Waldspurger's invertibility theorem spurred my interest for improving the state of the art in audio texture synthesis. Nevertheless, one important drawback of the scattering transform—in its original, purely temporal definition—is that it does not include the notions of relativity of pitch or relativity of tempo. Instead, each wavelet modulus layer decomposes all paths $p \to (\tau_{\alpha}, \psi_{\alpha})$ asynchronously. It was after personal communications with Shihab Shamma that we realized the crucial importance of accounting for joint modulations in time and frequency; or, said in algebraic terms, for elastic displacements over the affine Weyl-Heisenberg group on the primary auditory cortex. He also designed a multiresolution analysis scheme for time-frequency scattering, in the fashion of Mallat’s discrete wavelet transform algorithm and Simoncelli’s steerable pyramid. This scheme allowed interpreting the time-frequency scattering transform as the response of a deep convolutional neural network whose depth grows logarithmically with receptive field size. On my part, I wrote down the production rules of the time-frequency scattering coefficients into a multivariable framework. Upon advice from Joan Bruna, Assistant Professor of Computer Science and Data Science at New York University I chose to implement the stain by defining a prescription—starting from a random initial guess—usually Brownian motion noise—this procedure adds a corrective term to the signal at every iteration, so that its time-frequency scattering coefficients match those of a predefined texture target. Incidentally, it is also by means of stochastic gradient descent that most of the algorithms that are known today, albeit somewhat improperly, as “backpropagation.”

C.1 On Time-frequency Scattering and Computer Music:

wherein the multidimensional $\lambda$ encapsulates log-wavelengths $\gamma_{\lambda} \in \mathbb{R}$, particle spins $s \in \mathbb{T}$, and the infix operator $\cdot$ denotes list construction (“cons”) in the ML family of programming languages. This conceptual jump from purely temporal scattering to time-frequency scattering eventually turned out to be fruitful, but difficult: because wavelengths $\gamma_{\lambda}$ at one layer of the network (e.g. pitch $\gamma_1$ or tempo $\gamma_2$) may take over the roles of spatial variables $v_{\lambda}$ in a deep network, keeping track of all cross-dependencies between variables appealed for a more systematic resort to recursion in our numerical applications.

Andén and I studied the above definition in complementary ways. He used the principle of the stationary phase to confirm that time-frequency scattering characterizes the chirp rates of ripple stimuli, analogously to STRF in the primary auditory cortex. He also designed a multiresolution analysis scheme for time-frequency scattering, in the fashion of Mallat’s discrete wavelet transform algorithm and Simoncelli’s steerable pyramid. This scheme allowed interpreting the time-frequency scattering transform as the response of a deep convolutional neural network whose depth grows logarithmically with receptive field size. On my part, I wrote down the production rules of the time-frequency scattering coefficients into a multivariable framework. Upon advice from Joan Bruna, Assistant Professor of Computer Science and Data Science at New York University I chose to implement the stain by defining a prescription—starting from a random initial guess—usually Brownian motion noise—this procedure adds a corrective term to the signal at every iteration, so that its time-frequency scattering coefficients match those of a predefined texture target. Incidentally, it is also by means of stochastic gradient descent that most of the algorithms that are known today, albeit somewhat improperly, as “backpropagation.”
Aside from this technical distinction, audio texture synthesis from scattering coefficients is quite comparable to the training of a deep neural network. In both cases, the system produces uninformative output for a finite time and then, after being exposed to some real-world data, adjusts its own predictions by trial and error, until converging to a highly articulate statistical fit. For Joakim Andén and myself, refactoring the source code of the software library for scattering transforms so that it could allow for multi-threaded architectures and gradient backpropagation, was a steady effort of almost two years, with many emotional ups and downs—as is often the case in scientific research. By the end of 2015, we had a working implementation and presented it at the IEEE conference on Machine Learning for Signal Processing (MLSP) in Boston. Our paper boiled down to three claims: first, time-frequency scattering is more mathematically interpretable than other auditory representations, whether engineered or learned; secondly, on some tasks for which the availability of annotated data is limited (e.g. musical instrument recognition), it actually outperforms deep learning classifiers; and thirdly, it allows for the reconstruction of chips in audio textures, such as bird vocalizations, with satisfying perceptual similarity to the target. Yet, the section on signal re-synthesis was purely meant as an illustration of the capabilities and limitations of time-frequency scattering, as compared to other auditory representations. Never in the research agenda of my PhD did I anticipate that time-frequency scattering could one day prove to be useful to contemporary music creation.

Florian Hecker wrote to me for the first time in the spring of 2016. He had heard of time-frequency scattering through our collaborative work with Jean-Claude Risset, Associate Professor of Computer Science at KTH Royal Institute of Technology, Stockholm, and wanted to use it as a software for texture-related sound synthesis with wavelet features. When we first ran time-frequency scattering on his piece Modulator (2014), I was pleased to find that it performed quite well in terms of perceptual similarity, while conserving computational power in the scattering network. Contrary to other STRF-inspired software, the time-frequency scattering library was using a multiresolution pyramid to spare unnecessary computations in the lower frequencies; moreover, the wavelet factorization in Equation 3 allows for the vectorization of array operations and transforms (FFTs) to speed up convolutions. These technical improvements (although leaving the gist of the algorithm essentially unchanged) noticeably transformed the compositional workflow by allowing rapid prototyping of ideas. Because running one iteration of stochastic gradient descent now was about as long as the target sound clip, it became possible to listen to ideas. Because running one iteration of stochastic gradient descent now streamlined the compositional workflow by allowing rapid prototyping of ideas (although leaving the gist of the algorithm essentially unchanged) noticeably transforms (FFTs) to speed up convolutions. These technical improvements (although leaving the gist of the algorithm essentially unchanged) noticeably transformed the compositional workflow by allowing rapid prototyping of ideas. Because running one iteration of stochastic gradient descent now streamlined the compositional workflow by allowing rapid prototyping of ideas.

in the lower frequencies; moreover, the wavelet factorization in Equation 3 allows for the vectorization of array operations and transforms (FFTs) to speed up convolutions. These technical improvements (although leaving the gist of the algorithm essentially unchanged) noticeably transformed the compositional workflow by allowing rapid prototyping of ideas. Because running one iteration of stochastic gradient descent now streamlined the compositional workflow by allowing rapid prototyping of ideas.

One of the pieces we aimed at was the composition for the percussion work of the percussionist Adam Stone, who used the scattering network to create novel textures that could be used in the percussion work. This was an exciting project, as we were able to use the software to create new textures that could be used in the real world. The results were very promising, and we are now looking forward to working with other artists to explore the potential of this new technology.

In conclusion, time-frequency scattering is a powerful tool for audio texture synthesis. It allows for the creation of complex textures with high perceptual similarity to the original signal, and it is computationally efficient, allowing for real-time processing. This technology has the potential to revolutionize the way we think about audio textures and could be used in a wide range of applications, from music composition to sound design.
our other forms of communication, whether spoken, written, or via signs and therefore certainly not of this publication as an ersatz of post-serialist musical score. Rather, and despite the utter ineffability of music, it is possible to shed light upon our shared faculty of recursion, supplemented by perceptual quantization and tabular organization; of which musical notation is a mere by-product.

Far from any neo-numerological considerations, what is, in my mind, the intimate raison d'être of this publication, is that it helps us listeners understand two compositional prospects, and wraps them into one: the will to expand the scope of the potentially audible, by seeking for more and more complexity in the parametrization of sound synthesis; and the desire to delve deeper into what has been heard, by shifting the auditory focus onto previously unnoticed details. Music is, therefore, a two-fold ritual of anticipation. Like the composer, it is in the liminality of finite speeds that the faun shall dwell and thrive.

FAVN – Scattering to Text
Movement I