Fourier at the heart of computer music: From harmonic sounds to texture

Vincent Lostanlen, Joakim Andén, Mathieu Lagrange

Fourier at the heart of computer music: from harmonic sounds to texture

Vincent Lostanlen [a], Joakim Andén [b], Mathieu Lagrange [c]
[a] *Music and Audio Research Lab, New York University, New York, NY, USA*
[b] *Center for Computational Mathematics, Flatiron Institute, New York, NY, USA*
[c] *LS2N, CNRS, École Centrale de Nantes, Nantes, France*

Abstract

Beyond the scope of thermal conduction, Joseph Fourier's treatise on the *Analytical Theory of Heat* (1822) profoundly altered our understanding of acoustic waves. It posits that any function of unit period can be decomposed into a sum of sinusoids, whose respective contribution represents some essential property of the underlying periodic phenomenon. In acoustics, such a decomposition reveals the resonant modes of a freely vibrating string. The introduction of Fourier series thus opened new research avenues on the modeling of musical timbre—a topic which was to become of crucial importance in the 1960s with the advent of computer-generated sounds. This article proposes to revisit the scientific legacy of Joseph Fourier through the lens of computer music research. We first discuss how the Fourier series marked a paradigm shift in our understanding of acoustics, supplanting the theory of consonance of harmonics in the Pythagorean monochord. Then, we highlight the utility of Fourier's paradigm via three practical problems in analysis–synthesis: the imitation of musical instruments, frequency transposition, and the generation of audio textures. Interestingly, each of these problems involves a different perspective on time–frequency duality, and stimulates a multidisciplinary interplay between research and creation that is still ongoing. *To cite this article: V. Lostanlen, J. Andén, M. Lagrange, C. R. Physique X (2019).*

Résumé

Fourier au cœur de la musique par ordinateur : des sons harmoniques à la texture. Au-delà de son apport théorique dans le domaine de la conduction thermique, le mémoire de Joseph Fourier sur la *Théorie analytique de la chaleur* (1822) a révolutionné notre conception des ondes sonores. Ce mémoire affirme que toute fonction de période unitaire se décompose en une série de sinusoïdes, chacune représentant une propriété essentielle du phénomène périodique étudié. Dans l'acoustique, cette décomposition révèle les modes de résonance d'une corde vibrante. Ainsi, l'introduction des séries de Fourier a ouvert de nouveaux horizons en matière de modélisation du timbre musical, un sujet qui prendra une importance cruciale à partir des années 1960, avec les débuts de la musique par ordinateur. Cet article propose de thématiser l'œuvre de Joseph Fourier à la lumière de ses implications en recherche musicale. Nous retraçons d'abord le changement de paradigme que les séries de Fourier ont opéré en acoustique, supplantant un mode de pensée fondé sur les consonances du monocorde pythagoricien. Par la suite, nous soulignons l'intérêt du paradigme de Fourier à travers trois problèmes pratiques en analyse-synthèse : l'imitation d'instruments de musique, la transposition fréquentielle, et la génération de textures sonores. Chacun de ses trois problèmes convoque une perspective différente sur la dualité temps-fréquence, et suscite un dialogue multidisciplinaire entre recherche et création qui est toujours d'actualité. *Pour citer cet article : V. Lostanlen, J. Andén, M. Lagrange, C. R. Physique X (2019).*

Key words: Fourier analysis ; Computer music ; Audio signal processing

Mots-clés : analyse de Fourier ; musique par ordinateur ; traitement du signal audionumérique

1. Introduction

"Can the numbers with which a computer deals be converted in sounds the ear can hear?" In a 1963 article entitled *The Digital Computer as a Musical Instrument*, Bell Labs engineer Max Mathews raises this visionary question, which he immediately answers in the affirmative [32]. He argues that, indeed, the refinement of digital-to-analog conversion enables the composition of music under the form of a discrete-time sequence of amplitude values. After converting this sequence into a continuous-time electrical voltage signal, the computer would, by electromagnetic induction of some loudspeaker, emit a wave akin to those produced by conventional instruments.

Although the protocol described above may seem banal to the modern reader, it is important to stress the historical disruption that it represented, at the time, with respect to earlier technologies. On one hand, tape machines had an excellent fidelity in terms of playing back pre-recorded material, but offered little flexibility for further manipulation: variations in tape speed, for example, would typically affect both tempo and pitch proportionally. On the other hand, analog oscillators had multiple knobs and sliders for fine parametric control, but lacked the practical ability to approximate a diverse range of real-world sounds. Mathews was aware of this conundrum, and intended to demonstrate that computers could, in a near future, achieve a better tradeoff between expressivity and control than any other technology available at that time.

The prospect of using a computer to render sounds came with an important caveat. By virtue of the Nyquist–Shannon sampling theorem, encoding a continuous wave of bandwidth B without loss of information requires $2B$ discrete samples per second [47]. Given that the typical auditory system of humans has a bandwidth of about $B = 20\,\mathrm{kHz}$, this number amounts to around 40000 samples per second, that is, millions of samples for a musical piece of a few minutes. Appealing as it may seem to retain thorough control on the temporal evolution of the piece—down to microsecond time scales—the task of independently adjusting the amplitude value of each sample appeared, for the musician, to be a Sisyphean one.

"The numbers-to-sound conversion," Mathews points out, "is useless musically unless a suitable program can be devised for computing the samples from a single set of parameters." The presence of a central processing unit (CPU) within the digital computer, in addition to storage and actuation components, alleviates the design of sounds in the time domain. Consequently, the development of computer music requires a skillset at the intersection of acoustics, signal processing, and computer science. The combination of these skills was championed by Mathews and collaborators, notably John Chowning and Jean-Claude Risset, and further promoted by many others.

This article proposes an abridged history of mathematical models for the production and perception of musical sounds. Throughout this ongoing quest for a musical *lingua franca* between musicians and scientists, the legacy of Joseph Fourier plays a pivotal role in at least three aspects. First, representing a periodic function by its Fourier series neatly disentangles its fundamental frequency, a continuous and one-dimensional parameter, from its spectral envelope, a potentially infinite sequence of complex-valued numbers. Secondly, the fast Fourier transform algorithm (FFT) allows efficient convolutions between any two discrete-time signals, even if one of them has an infinite impulse response. Thirdly, in the study of audio textures, computing the Fourier transform of the autocovariance function reveals the power spectral density of the underlying stationary process. None of these research topics were properly formulated, let alone addressed, by the law of arithmetic resonance between the harmonics of a vibrating string, which had remained hegemonic from Ancient Greece to the Enlightenment, under the name of monochord. Nevertheless, in each of them, the resort to Fourier theory nicely bridges the gap between numerical and perceptual representations of musical sound.

A comprehensive review of modern techniques in Fourier-based audio processing lies beyond the scope of this article; for this purpose, we refer the reader to [53]. Rather, we choose to restrict our narrative to five basic methods: the additive sinusoidal model, the short-term Fourier transform (STFT), the power cepstrum, frequency modulation (FM) synthesis, and the phase vocoder. Their integration into digital audio workstations has shaped the history of post-war art music, and also made its way into pop music.

Email addresses: vincent.lostanlen@nyu.edu (Vincent Lostanlen), janden@flatironinstitute.org (Joakim And),

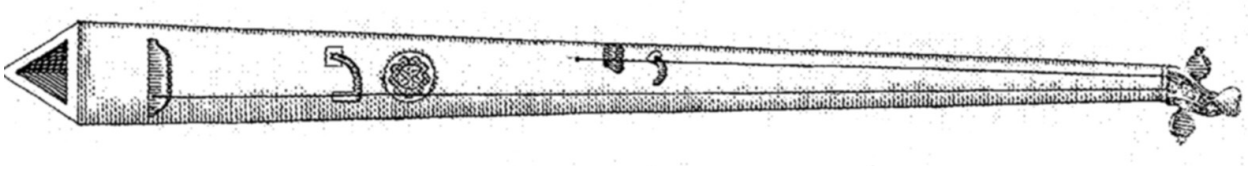

Figure 1. Drawing of a monochord by Athanasius Kircher, published in his treatise *Musurgia Universalis* (1650).

In this article, our primary ambition is to show that the duality between the time domain and the Fourier domain brings conceptual coherence to the profusion of computer music systems that followed Mathews's breakthrough. Furthermore, we show that, although the paradigm of Fourier *stricto sensu* is insufficient to synthesize aperiodic sources, this paradigm can be extended to a greater level of generality by resorting to a multiresolution scheme. Over the past two decades, researchers in audio signal processing have developed several alternatives to the short-term Fourier transform to improve the analysis–synthesis of natural audio textures, such as field recordings of wildlife. This article showcases one such alternative, joint time–frequency scattering, which is based on the extraction of spectrotemporal modulations in the wavelet scalogram. In doing so, our goal is not to offer an definitive account of the state of the art in audio texture synthesis, but to discuss the relevance of Fourier's paradigm in currently active endeavors of computer music research.

Section 2 retraces the history of reductionist models of acoustic waves, from the Pythagorean monochord to Fourier's *Analytical Theory of Heat*, followed by their digital implementation in the 1960s. Section 3 presents some advanced capabilities of Fourier-based techniques for the analysis and synthesis of nonstationary sounds. Lastly, Section 4 extends the Fourier paradigm to a deep multiresolution framework by introducing time–frequency scattering and its application to audio texture synthesis.

2. Separable Fourier series synthesis

The first questions on the nature of sound came from a physical perspective: how do physical objects, such as vibrating strings, result in audible sounds? It was not until the introduction of Fourier analysis that a finer understanding of this question was achieved. In particular, the notion of a Fourier series to describe a periodic function proved fruitful in, first, analyzing sounds, but eventually, also in synthesizing them. This latter application was introduced in the work of Max Mathews and his MUSIC software, which used a Fourier series modulated by a temporal envelope to create some of the first computer-generated musical sounds.

2.1. *Pythagorean monochord*

Finding an adequate decomposition of music into atomic entities is a problem that dates back, at least, to Ancient Greece and the institutionalization of the mathematical proof. The Pythagorean monochord (*kanon*), a one-string zither comprising a movable bridge and a graduated rule, finds its oldest known written trace in Euclid's *Division of the Canon*. One may adjust the length of the vibrating part of the string by moving the bridge at a specific distance from the monochord nut. Although rudimentary, and deliberately conceived as a thought experiment, the introduction of the monochord established two essential principles of musical acoustics. The first is descriptive: all other things being equal, the musical interval between two strings is determined by the ratio of their vibrating lengths. The second, in contrast, is prescriptive: intervals whose irreducible length ratios have a small integer denominator elicit a sensation of greater consonance.

Once these two principles are taken for granted as axioms, the quest for a consonant temperament [20] amounts to the following arithmetic problem: what are two finite integer sequences $p_1, \dots, p_N$ and $q_1, \dots, q_N$ such that all pairs of cross-terms $p_n q_{n'}$ and $p_{n'} q_n$ have a small common product? For $N = 3$, a solution is given by the perfect fourth (4/3), the perfect fifth (3/2), and the octave (2/1). These three intervals are at the core of Ptolemy's theory of musical tuning, where they are known as epimoric ratios. Diatonic scales ($N = 7$), however, necessarily incur harsh dissonance in at least one pair of tones, as defined by the above Pythagorean tuning axioms. The topic of mitigating this dissonance and seeking a "well-tempered"

scale which proceeds from the method of dividing the string into subparts of rational length has sparked a controversy that lasted across two millennia. Numerous treatises, from Aristoxen of Tarent and Boethius to René Descartes, have addressed this controversy from a multidisciplinary standpoint, combining mathematics, music, and philosophy [3].

Despite never seeing the light of stage, the monochord acted as a shared paradigm for scholars, composers, and manufacturers. Indeed, it promoted, albeit somewhat inchoately, two fundamental notions in physics: modal resonance and wave superposition. Modal resonance states that a fixed body can, in its stationary regime of vibration, be understood by a countable number of elementary *eigenmodes* [21]. Wave superposition states that, in nature, vibrating bodies do not tune in exclusively to one eigenmode or another, but oscillate according to some mixture thereof [36]. In the parlance of dynamical systems, the monochord is a prime example of conceptual disentanglement between shape (boundary conditions) and state (initial conditions).

2.2. *The Fourier revolution*

The age of the Enlightenment marks the culmination of the monochord paradigm, up to the point of revealing its intrinsic deficiencies. In the *Encyclopédie*, begun in 1749, Jean-Jacques Rousseau coined the term *timbre* to refer to acoustic qualia from dull to bright and from sour to sweet. He pointed out that the perception of timbre allows the listener to recognize the identity of instruments, even when those instruments play notes of identical intensity and pitch. Understanding timbre requires investigating the physical interaction between shape and state, and in particular, how the playing technique of a note (e.g., plucked or bowed) affects the relative magnitudes of all superposed eigenmodes through time. Yet, the monochord paradigm does not offer any experimental protocol for gauging these relative magnitudes, and is thus inadequate for discussing the physical bases of musical timbre perception. Similarly, in his treatise on *Harmony Reduced to its Natural Principles* (1722), composer Jean-Philippe Rameau concluded that consonances on the monochord did not suffice for explaining tonal harmony in the common practice period. Both Rameau and Rousseau communicated with mathematician Jean Le Rond d'Alembert about the need to overhaul the acoustical framework inherited from Ancient Greece. This led d'Alembert to write a treatise on the *Elements of music* (1752), which acknowledged the obsolescence of the monochord and proposed some future directions for studying music. Although d'Alembert had discovered, shortly earlier (in 1749), the partial differential equation governing the motion of vibrating strings, the connection between the initial state of this equation and the superposition of eigenmodes remained unclear for seven more decades, until the publication of Joseph Fourier's *Analytical Theory of Heat* (1822).

Although not directly related to music, Fourier's treatise represented a turning point in the understanding of the d'Alembert wave equation [16]. The treatise proposes to isolate the question of heat propagation from the broader study of heat, and notably from its chemical and dynamical aspects [17]. In doing so, it established an autonomous branch of mathematical physics, later known as harmonic analysis. Fourier was certainly not the first to study trigonometric series of the form $\sum_p a_p \cos(2\pi p\xi t + \varphi_p)$ for arbitrary sequences of amplitudes a_p and phases φ_p: these series are found, for instance, in the writings of d'Alembert, as well as Euler and Bernoulli. However, the innovation of his treatise lies in the claim that such trigonometric series are universal, in the sense that they can be applied not only to construct eigenmodes of the Laplacian operator, but also to represent the initial condition of any one-dimensional heat conduction problem.

The Fourier series representation is more than an intermediate computational step for disentangling spatial and temporal variables in one partial differential equation. According to epistemologist Alain Herreman [19], it inaugurates the duality between the set of "arbitrary curved lines" (in French, « lignes courbes tracées arbitrairement » [16]) and the set of analytical expressions based on trigonometric series. Whereas the former set is taken for granted as proceeding from the real world, the second set is written in the language of calculus and is deliberately kept apart from any physical instantiation. Semiotically speaking, Fourier argues that the latter is in a relation of *incommensurable conformity* (in the words of Herreman) with the former. Once translated to acoustics, this statement means that, although Fourier series are not a necessary consequence of observing the natural vibration of strings (incommensurability), they have the merit of potentially representing any waveform (conformity). By unicity of the Fourier series, the identification of

amplitudes a_p and phases φ_p relies on complex-valued integrals of the form

$$a_p \exp(\mathrm{i}\varphi_p) = \xi \int_0^{1/\xi} \boldsymbol{x}(t) \exp(-2\pi \mathrm{i} p \xi t)\, \mathrm{d}t \quad \forall p \in \mathbb{N}^* \qquad \text{with} \qquad \boldsymbol{x}(t) = \sum_{p=1}^{+\infty} a_p \cos(2\pi p \xi t + \varphi_p). \tag{1}$$

The brilliant intuition of Fourier is that, in practice, the function $\boldsymbol{x}$ may encode the initial relative temperature in a metal rod (in which case t is a one-dimensional spatial variable), but also the initial displacement of a vibrating string. "If one applies these principles to the question of the motion of vibrating strings," Fourier writes, "one will resolve the difficulties that their analysis by Daniel Bernoulli had raised in the past." He continues: "This question differs a lot from that of the distribution of heat; but the two theories have points in common; because both are founded on the analysis of partial derivatives."

2.3. *Modeling timbre with the spectral envelope*

Over the 142 years between Fourier's *Analytical Theory of Heat* and Mathews's *Digital Computer as a Musical Instrument*, multiple scholars progressively acknowledged the importance of timbre in music perception, and the convenience of thinking about sound in terms of both time and frequency. In this regard, two historical milestones are the publication of a *Treatise on Instrumentation* (1844) by composer Hector Berlioz, and the *Sensations of Tone* (1863) by physicist Hermann von Helmholtz. By the time digital computers became commonplace, the resort to Fourier series appeared to Mathews as self-evident. The major appeal behind adopting Fourier series as a signal model for $\boldsymbol{x}$ is that they single out fundamental frequency ξ as a continuous, one-dimensional parameter of sound perception, whose effect spans over all discrete-time samples t. Moreover, Fourier series allow, up to some extent, to address the question of timbre modeling by parametrizing amplitude coefficients a_p independently of the choice of fundamental frequency ξ. The coefficients a_p correspond to the values of the spectral envelope of the source at evenly spaced frequencies $p\xi$, wherein the integer p corresponds to the index of one of the sinusoids, denoted as harmonic partial. This spectral envelope, in turn, relates to the shape and material of the instrument, as well as the gesture of the performer. Therefore, although the identification of a source goes beyond the perception of its spectral envelope, the values of a_p certainly have a central role in timbre perception. For the composer, the choices of ξ and a_p resemble choices in harmony and orchestration, two well-established attributes of artistic style in Western art music. Such a disentanglement is crucial to ensure that a computer program can parse, process, and render any musical score.

2.4. *The MUSIC software environment*

The software that Max Mathews built in 1963, simply dubbed MUSIC, relied heavily on Fourier series to define the timbre of virtual instruments. The layout of early computer music programs closely resembled electronic circuits of oscillators and modulators, with different parameters for duration, loudness, and amplitude, such as those already in use at the time by composers in analog-based electronic music. However, contrary to a tangible analog circuit, the complexity of a computer music program is not bound by constraints of human operability, but only by the speed of the central processing unit. Whereas soldering several hundreds of oscillators together would be a cumbersome task for the composer, instantiating these oscillators in a virtual environment considerably alleviates this task, as it delegates the bulk of parametrization to the computer. The falling costs of hardware from the 1960s onwards allowed Mathews's MUSIC programs (notably its 1966 version, the widely acclaimed MUSIC-V [33]) to gain in sophistication and detail. These programs were distributed with a collection of preset values for the sequence of Fourier coefficients a_p. Likewise, predefined amplitude functions $t \mapsto \boldsymbol{\alpha}(t)$ allowed the user to model the temporal profile of attack, decay, sustain, and release (ADSR) of these notes, as a piecewise linear function comprising four segments. On the part of the composer, the degrees of creative freedom are then: fundamental frequency ξ, temporal profile $\boldsymbol{\alpha}$, and onset time τ. Under this framework, for a virtual instrument with P harmonic partials playing a piece with N notes, the computer rendition of the musical piece is given by

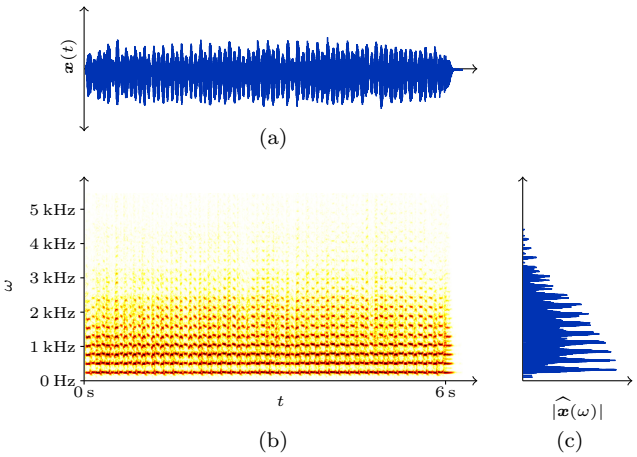

Figure 2. (a) A recording $\boldsymbol{x}$ of a trumpet playing a trill with pitch C4. (b) The spectrogram of $\boldsymbol{x}$. (c) The Fourier transform magnitude of $\boldsymbol{x}$.

$$\boldsymbol{x}(t) = \sum_{n=1}^{N} \boldsymbol{\alpha}_n(t - \tau_n) \sum_{p=1}^{P} a_p \cos(2\pi p \xi_n t + \varphi_p). \tag{2}$$

In order to match the tone quality of familiar instruments, Max Mathews and his collaborator Joan Miller implemented Fourier-based synthesis in MUSIC. While inferring preset values of a_p by reproducing the Fourier series of external data did not yield new sounds per se, it provided a convenient starting point for timbral exploration. Because the number of partials P with non-negligible amplitudes a_p was typically of the order of 20, such adjustment was humanly tractable—as opposed to the sample-wise description of the waveform $\boldsymbol{x}(t)$ in the time domain, which would typically involve a hundred samples per period or more. In addition, a single preset of values $a_1, \ldots, a_p$ would suffice to span a wide tessitura of pitch values ξ_n, for a given instrument.

To illustrate the frequency content of a sound $\boldsymbol{x}$, let us consider its *short-time Fourier transform* (STFT):

$$\mathrm{STFT}(\boldsymbol{x})(t, \omega) = \int_{-\infty}^{\infty} \boldsymbol{x}(u)\boldsymbol{g}(u - t) \exp(-2\pi\, \mathrm{i}\omega u)\, \mathrm{d}u \tag{3}$$

where $\boldsymbol{g}(u)$ is the window function associated with the STFT. This representation of $\boldsymbol{x}$ lets us examine its frequency content when restricted to the window $\boldsymbol{g}$ centered at time t. Taking its modulus, we obtain the *spectrogram* $|\mathrm{STFT}(\boldsymbol{x})(t, \omega)|$.

Figure 2(a) shows the six-second waveform of a trumpet playing the note C4 with a trill. The corresponding spectrogram, with a Hann window of length 46 ms, is shown in Figure 2(b). Finally, the Fourier transform magnitude of the whole signal (i.e., not localized using a window) is shown in Figure 2(c). Here, the regular harmonic structure induced by the constant pitch is readily seen in both the spectrogram and Fourier transform magnitude. The spectrogram, furthermore, reveals the dynamics of the signal at different frequencies—each traces out a distinct envelope. There are also slight changes in the pitch, as revealed by the slight oscillation of the partial contours. As we shall see, both of these phenomena also contribute greatly to the perception of a sound.

2.5. *Parametrizing the spectral envelope*

The P values $a_1, \ldots, a_P$ can, in turn, be encoded by even fewer parameters, which would encompass the overall shape of the spectral envelope. For example, the amplitude ratio between some odd-numbered partial a_{2p+1} and its odd neighbor a_{2p} has an interpretable physical meaning, in terms of boundary conditions of the d'Alembert wave equation. The bore of a transverse flute is open on both ends: as a result, the odd-to-even energy ratio is of the order of 1. The bore of a clarinet, on the other hand, is open on the lower end and

closed on the other: the presence of the reed weakens even-numbered partials, thus increasing the odd-to-even energy ratio [15]. Such an insight into the physical underpinnings of sound production, derived from a Fourier decomposition, allows the composer to interpolate between two well-known types of musical timbre (flute and clarinet) by controlling a single continuous parameter.

Another example of a one-dimensional feature for musical timbre that has a physical interpretability is the rate of decay of Fourier amplitude coefficients, also known as spectral slope. On one hand, solving the d'Alembert equation with zero initial velocity and nonzero initial displacement at the midpoint of the string yields a Fourier series whose amplitude terms a_p are proportional to $1/p^2$. On the other hand, solving the same equation with zero initial displacement and nonzero initial velocity leads, all other things being equal, to Fourier coefficients of greater magnitudes: for the first values of p, magnitudes are almost proportional to $1/p$. This discrepancy in decay is reflected in the time domain by a discrepancy in waveform shape, from triangular to square. Mathematically speaking, this duality between regularity in the time domain and rate of decay in the Fourier domain is at the heart of the notion of Hölder continuity and Sobolev spaces [49]. A slower decay incurs a relative increase in the energy of the signal $\boldsymbol{x}(t)$ at high frequencies, which is perceived as a sensation of acoustic brightness [34]. In musical acoustics, the case of zero initial velocity and nonzero initial displacement corresponds to an instrument with plucked strings (e.g., a guitar or harpsichord), whereas the case of nonzero initial velocity and zero initial displacement corresponds to an instrument with hammered strings (e.g., a piano). In a computer music environment, parametrizing the Fourier magnitudes as proportional to $1/p^s$ for arbitrary s yields a seamless interface for interpolating acoustic brightness between the typical Fourier series of a plucked string ($s \to 1$) and that of a hammered string ($s \to 2$).

Over and above the physical interpretability of the two aforementioned features, they also account for independent aspects of the spectral envelope. Indeed, the odd-to-even energy ratio and the rate of decay of Fourier coefficients characterize local and global contrast between partials, respectively.

2.6. *Power cepstrum*

In between these two extremes, it is possible to define a family of descriptors that compute contrast over the entire Fourier spectrum according to various scales, ranging from coarse (e.g., the rate of decay) to fine (e.g., the odd-to-even energy ratio). In a 1963 paper by Bruce Bogert, Michael Healy, and John Tukey [6], the authors transform a_p by applying a pointwise logarithm followed by the Fourier series summation formula:

$$\boldsymbol{c}(k) = \sum_{p=1}^{+\infty} \log(a_p) \exp\left(2\pi \mathrm{i} \frac{pk}{P}\right). \tag{4}$$

The Fourier sum operates as though the discrete frequency variable p were a temporal dimension, and yields a continuous variable k which is physically homogeneous to a time span. Therefore, Tukey proposed to refer to $\boldsymbol{c}(k)$ as a power *cepstrum*, an anagram on the word "spectrum." Likewise, through an anagram on "frequency," the variable k is known as a *quefrency*, and is expressed in Hz^{-1}, i.e., in seconds. At high quefrency $k = P/2$, the complex exponential in Equation 4 boils down to $(-1)^k$, and the corresponding cepstral coefficient approximates the average logarithm of the even-to-odd energy ratio. Conversely, at low quefrency $k = 1/2$, the modulus of $\boldsymbol{c}(k)$ is proportional to the exponent s in the decay of Fourier coefficients.

3. Analysis–synthesis in the time–frequency domain

The resort to spectral shape descriptors and to cepstral coefficients is central to the perceptual modeling of timbre similarity, and, to this day, finds applications in both analysis and of synthesis of audio signals. However, these tools are inadequate for nonstationary sounds; that is, signals whose spectral envelope varies through time, either by effect of gestural expressivity or due to nonlinearities of the acoustic resonator with respect to input amplitude. Two examples of such nonstationarities are the vibrato in the case of the violin and the crescendo in the case of the trumpet. The imitation of these two instruments, conducted by Mathews and Risset, respectively, was not conceived as an end in itself, but rather as a well-defined test bed for the

development of new synthesis models, ultimately leading to computer-generated sounds that were never heard before. This section describes how to extend the fundamental principles of Fourier analysis–synthesis in order to improve its range of applicability for computer music.

3.1. *Bridging the gap between expressivity and control*

Before the democratization of computer music programs, Mathews's idea of using the digital computer as a musical instrument was met with skepticism and disbelief. This was due, in part, to the widespread impression that computers were best suited for rule-based reasoning, not for creative expression. With the important exception of Ada Lovelace, who, in 1842, had speculated upon the potential use of Charles Babbage's *Analytical Engine* for writing music, few had anticipated that the temporal unfolding of a musical piece could be obtained as the result of a computation. Instead, up until the 1960s, scientific research at the intersection between music and technology was polarized around two paradigms: *musique concrète* and *elektronische Musik*. The former, championed by Pierre Schaeffer at the Studio d'Essai in Paris, was based on the recording and manipulation of magnetic tapes [44]. The latter, championed by Karlheinz Stockhausen at the Studio for Electronic Music in Cologne, was based on the modular association of analog oscillators [18].

In this context, Mathews faced the challenge of demonstrating that digital computers could strike a satisfying tradeoff between the versatility of *musique concrète* with the controllability of *elektronische Musik*. Therefore, he undertook the time-consuming task of improving the fidelity of presets in the MUSIC program, in particular for instruments that were not modeled well by Equation 2, that is, as a Fourier series multiplied by an ADSR envelope. One of the core findings of Schaeffer, indeed, had been to exhibit the inherently spectrotemporal nature of auditory perception. By manipulating the temporal envelope of pre-recorded sounds, he showed that the attack part of the note (the A in ADSR) had a decisive importance in the identifiability of a musical instrument. For example, fading the potentiometer during the attack part of a bell sound is enough to erase its percussive attributes and mutate it into an oboe-like tone [43]. The same is true of many other dynamic aspects of timbre perception, such as vibrato and tremolo: although their influence on cepstral coefficients is tiny, they convey a sensation of acoustic "vitality" which, when discarded, may leave a sensation of robotic affectlessness.

3.2. *Frequency-modulated Fourier series*

To simulate the typical vibrato of the violin, Mathews implemented frequency modulation in the synthesis model. By way of two user-defined parameters, the modulation rate ν and the modulation depth Δ, the audio signal is given by

$$\boldsymbol{x}(t) = \sum_{n=1}^{N} \boldsymbol{\alpha}_n(t-\tau_n) \sum_{p=1}^{P} a_p \cos\Big(p\Big(\xi_n t + \frac{\Delta}{\nu}\sin(\nu t)\Big)\Big). \tag{5}$$

The above equation musical notes of time-varying fundamental frequency $\xi_n + \Delta\cos(\nu t)$. Incidentally, it found applications in computer music well beyond the modeling of vibrato. In 1967, John Chowning realized that setting the value of $\nu = \xi_1$, or even $\nu > \xi_1$, allowed to synthesize a surprisingly rich set of non-vibrating tones, yet with no more than three parameters: ξ_1, ν, and Δ [9]. Because its Fourier series is implicitly encoded into the composition of two sinusoidal functions, computing a musical note via Chowning's frequency modulation (FM) synthesis is considerably faster than via Mathews's additive synthesis. The invention of FM synthesis inaugurated the distribution of real-time, inexpensive digital music tools for pop music productions.

3.3. *Spectrotemporal modulation*

Besides vibrato, another musical effect which is beyond reach of the Fourier series summation model in Equation 1 is crescendo in brass instruments. Indeed, owing to nonlinear wave propagation along the bore, as well as variations in lip movement and viscothermal loss at the bell [4], the acoustic brightness of a brass

instrument is highly dependent on the sound level. Going back to the additive model of Equation 2, this implies that the temporal variations of global amplitude $\boldsymbol{\alpha}_n(t-\tau_n)$ should affect each partial p differently [8]. In 1964, composer Jean-Claude Risset, while on a visit to Bell Labs, implemented a generalized version of additive synthesis for MUSIC IV. In this version, the computer-generated waveform is

$$\boldsymbol{x}(t) = \sum_{n=1}^{N} \sum_{p=1}^{P} \boldsymbol{\alpha}_{p,n}(t-\tau_n) \cos(p\xi_n t), \tag{6}$$

where each function $\boldsymbol{\alpha}_{p,n}$ represents the temporal amplitude curve of partial p for note n. Risset employed this nonstationary model to synthesize some melodic elements in *Computer Suite for Little Boy* (1968) and *Mutations* (1969), two milestone pieces in the development of computer music.

Despite allowing for more expressiveness than Equation 2, Equation 6 suffers from a considerable increase in complexity. This is because it trades a separable model of temporal envelope $\boldsymbol{\alpha}_n(t-\tau_n)$ and spectral envelope a_p for a joint spectrotemporal model, in which the amplitude functions $\boldsymbol{\alpha}_{p,n}(t-\tau_n)$ do not necessarily factorize as a separable product of a temporal and a spectral envelope. With this transition, the resort to data-driven presets, rather than pure trial and error, became increasingly important. To compute these presets, Risset applied a family of P bandpass filters $\boldsymbol{\psi}_{p\xi_n}$ with center frequencies at $p\xi_n$ over pre-recorded notes $\boldsymbol{y_n}$ of known fundamental frequency ξ_n.

3.4. *Spectrogram analysis and re-synthesis*

One way to obtain the amplitude functions $\boldsymbol{\alpha}_{p,n}$ from each of these bandpass filters is to define them as $\boldsymbol{\alpha}_{p,n}(t) = |\boldsymbol{y_n} * \boldsymbol{\psi}_{p\xi_n}|(t)$, where the asterisk $*$ denotes the convolution operation

$$(\boldsymbol{x} * \boldsymbol{y})(t) = \int_{-\infty}^{+\infty} \boldsymbol{x}(\tau)\boldsymbol{y}(t-\tau)\,\mathrm{d}\tau \tag{7}$$

and the vertical bars ($|z| = \sqrt{z\bar{z}}$) denote complex modulus. This procedure was to herald a long-standing paradigm in computer music, known as *analysis–synthesis*. Here, the Fourier domain is neither a starting point nor an end point, but an intermediate step, in which computations can be expressed more naturally than in the time domain [42]. To produce an imitation of a trumpet, Risset began by analyzing a real-world signal $\boldsymbol{y_n}$ with an analog electronic device named the *sound spectrograph* [22], invented at Bell Labs in 1946. He then reduced the description of $\boldsymbol{y_n}$ to a few, slowly varying amplitude functions $\boldsymbol{\alpha}_{p,n}$. Lastly, he resynthesized the original signal by a procedure similar to Fourier series summation. Upon his return to France in 1965, Risset showcased his synthesized trumpet tones to the French Academy of Sciences [39].

In 1966, audio spectral analysis suddenly became less computationally demanding, thanks to the invention of the fast Fourier transform (FFT) by James Cooley and John Tukey. Indeed, the FFT brought the complexity of discrete-time convolutions between two signals of length L from $O(L^2)$ down to $O(L \log L)$. The discrete Fourier transform $\mathcal{F}(\boldsymbol{\psi})$ of some bandpass filter ψ of length L is defined as:

$$\mathcal{F}(\boldsymbol{\psi}) : \omega \in \{0, \ldots, L-1\} \longmapsto \widehat{\boldsymbol{\psi}}[\omega] = \sum_{\tau=0}^{L-1} \boldsymbol{\psi}(\tau) \exp\left(-2\pi \mathrm{i} \frac{\omega\tau}{L}\right). \tag{8}$$

The circular convolution theorem states that the application of the discrete Fourier transform converts a convolution product into an element-wise multiplication: $\mathcal{F}(\boldsymbol{x} * \boldsymbol{y}) = \mathcal{F}(\boldsymbol{x}) \times \mathcal{F}(\boldsymbol{y})$. Therefore, it is possible to efficiently measure all amplitude modulation functions by multiplication in the Fourier domain followed by pointwise complex modulus in the time domain: $\boldsymbol{\alpha}_{p,n}(t) = |\mathcal{F}^{-1}(\widehat{\boldsymbol{y}} \times \widehat{\boldsymbol{\psi}}_{p\xi_n})|(t)$. Furthermore, such an analysis may still be carried out without prior knowledge of the fundamental frequency ξ, and even for input signals $\boldsymbol{y}_n(t)$ with no discernible fundamental frequency at all. The use of band-pass filters ψ generalizes time–frequency models from a short-term Fourier series with a signal-dependent period ξ^{-1} to an STFT with a signal agnostic frame size L, under an assumption of local, wide-sense stationarity. This generalization is important for computer music, because it opens the possibility to perform analysis–synthesis on aperiodic, noisy, or even fractal sounds [45].

3.5. *Phase vocoder and applications*

The analysis–synthesis paradigm is the digital equivalent of the channel *vocoder*, invented by Homer Dudley at Bell Labs in 1940 [12]. Both paradigms, the analog and the digital one, rely on the Fourier transform in order to map signals onto the time–frequency domain. The digital formulation, however, offers finer control on $\boldsymbol{\alpha}_{p,n}$ and, notably, the possibility to perform frequency transposition and time stretching. This idea was proposed in 1966 by Bell Labs engineer James Flanagan, under the name of the *phase vocoder* [14]. Risset coined the terms of *sonic microsurgery* and *intimate transformations* to denote such operations of artificial time stretching and frequency transposition [38]. We refer to [24] for an introduction to the phase vocoder and to [5, 27] for an overview of the state of the art in the domain.

As long as the amount of time stretching is small enough, the computer-generated result remains perceptually realistic, and is hardly discernible from an actual recording at a new tempo. Thus, the phase vocoder has found applications in real-time score following [10], so as to adjust the tempo of a pre-recorded accompaniment onto the expressive tempo fluctuations of a human performer. However, for extreme values of time scale modification, phase vocoding produces sounds which are no longer realistic, but may still be interesting musically speaking. For example, composer Trevor Wishart and computer scientist Mark Dolson released *Vox 5* in 1986, a computer music piece in which a spoken voice stretches through time, progressively loses its recognizable features, and eerily mutates into non-speech sounds, such as a neighing horse or rumbling thunder [52].

Another application of the phase vocoder is real-time pitch adjustment. This technology has made its way into pop music under the registered trademark of "Auto-Tune." After tracking the pitch curve $\boldsymbol{\xi}_n(t)$ of each note n in a live audio stream, Auto-Tune replaces $\boldsymbol{\xi}_n(t)$ by its closest neighbor over some predefined musical scale. This replacement produces a singing voice that is perfectly in tune, often to the point of sounding uncanny. Up until today, some musicians have embraced this uncanniness and use the phase vocoder to build, in the word of artist Jace Clayton, a "duet between the electronic and the personal" [48].

4. Wavelet scattering of audio textures

From all of the above, it appears that the invention of the fast Fourier transform has allowed computer music researchers to move away from the rigid template of the harmonic series, and explore the design space of amplitude modulation (AM) as well as frequency modulation (FM). Indeed, the modeling of transients for specific instruments by Risset and Mathews eventually gave way to partial tracking [46] and phase vocoding, two complementary algorithms for adaptively stretching pitch contours in the time–frequency domain. However, the musical applicability of these algorithms is predicated on the assumption that the sonic material at hand consists of a finite sum of sinusoids, and that these sinusoids have slowly varying amplitudes and frequencies in comparison with some time scale. Although this assumption is often valid for many samples of tonal music, it falls short for audio textures, such as large orchestral clusters, drum rolls, and field recordings of wildlife.

In some cases, it remains possible to accommodate noisy components within a sinusoidal modeling framework by a procedure known as bandwidth enhancement [13]. However, the prospect of modeling aperiodic sources calls for a multiscale generalization of Fourier analysis–synthesis. Known as wavelet decomposition, this multiscale generalization is well suited for representing highly oscillatory signals, but also characterize transient phenomena, as evidenced by experiments in psychoacoustics [35]. This section presents a wavelet-based approach to audio texture synthesis. We ought to stress that, in this regard, wavelets are by no means hegemonic; we refer to [26] for a recent overview of the state of the art in sound texture modeling based on the STFT representation.

4.1. *Wavelet transform*

In 1971, shortly after his return from Bell Labs, Risset was invited by the University of Marseille to found a research team focused on computer music that, in 1978, became part of the Laboratory of Mechanics

and Acoustics of the CNRS. In Marseille, Risset met quantum physicist Alex Grossmann, thereby sparking a collaboration between Grossmann's Center for Theoretical Physics and his own laboratory. The year 1984 marks an acceleration of this collaboration, around the emergent topic of *wavelets*—a term coined by Grossmann to denote families of well-localized functions of constant shape and varying bandwidth. One example of such well-localized function is the Morlet wavelet, defined as

$$\boldsymbol{\psi}_\lambda(t) = \lambda \exp\left(-\frac{\lambda^2 t^2}{2Q^2}\right) \times (\exp(\mathrm{i}\lambda t) - \kappa), \tag{9}$$

where λ is the center frequency, Q is the quality factor, and κ is calibrated so $\boldsymbol{\psi}_\lambda(t)$ has zero mean. Note that, while this wavelet is complex-valued, it is possible, and often desirable, to define real-valued wavelets. The wavelet representation has many interesting theoretical properties that often make it better suited to the analysis of nonstationary signals than the STFT, which is based on a family of well-localized functions of constant bandwidth [30]. From a musical perspective, the most compelling of these properties is that the quality factor Q defines a just-noticeable difference in pitch perception around every λ, as an interval on a chromatic scale. For instance, setting $Q = 1$ leads to a relative bandwidth of about one octave, whereas $Q = 12$ leads to a bandwidth of about one semitone in twelve-tone equal temperament. In contrast, the choice of window in the STFT defines an absolute bandwidth in Hertz, regardless of the center frequency.

On one hand, the modulus of the STFT yields a spectrogram whose vertical axis grows linearly with frequency ω. On the other hand, the continuous wavelet transform yields a time–frequency representation

$$\mathbf{U_1}(\mathbf{x})(t, \log\lambda) = |\boldsymbol{x} * \boldsymbol{\psi}_\lambda|(t), \tag{10}$$

whose vertical axis grows logarithmically with frequency λ. The representation $\mathbf{U_1}(\boldsymbol{x})$ is known as the *scalogram* of $\boldsymbol{x}$. Although a change in the fundamental frequency ξ of a musical note would result in a scaling of the vertical axis of the STFT, its effect on the wavelet scalogram is a simple translation, due to this logarithmic mapping. Therefore, the wavelet scalogram is particularly well suited to the implementation of Risset's *intimate transformations*, such as time stretching or frequency transposition [37].

In 1985, Risset and Grossmann initiated a special interest group around the topic of wavelets that included mathematicians Yves Meyer and Ingrid Daubechies, geophysicist Jean Morlet, quantum physicist Thierry Paul, signal processing researchers Daniel Arfib and Richard Kronland-Martinet, and many others. Within this group, Kronland-Martinet, Grossmann, and Morlet implemented the continuous wavelet transform in the SYTER software environment (in French, *Système temps réel*) for audio analysis and re-synthesis [23]. This software had a profound impact on computer music research. Indeed, the visual layout of the wavelet scalogram $\mathbf{U_1}(\boldsymbol{x})$ mirrors the representation of musical notes on a score, with time along the horizontal and pitch along the vertical axis. This layout is thus particularly intuitive for a classically trained musician.

4.2. *Phase retrieval*

Under relatively mild assumptions on $\boldsymbol{\psi}$, the wavelet transform is an invertible operator, with a stable, closed-form inverse [30, Theorem 4.4]. The same cannot be said of the wavelet scalogram operator $\mathbf{U_1}$: because the application of pointwise complex modulus incurs a loss of phase information, recovering the signal $\boldsymbol{x}$ from its scalogram $\mathbf{U_1}(\boldsymbol{x})$ is far from trivial. Yet, this nonlinear operation of complex modulus has the advantage of demodulating locally periodic oscillations in $\boldsymbol{x}$, thus allowing to model the typical variations in $\mathbf{U_1}(\boldsymbol{x})$ with smoother control parameters. As a result, it converts a complex-valued, rapidly changing function $(\boldsymbol{x} * \boldsymbol{\psi})(t, \log\lambda)$ into a real-valued, nonnegative, slowly varying function $|\boldsymbol{x} * \boldsymbol{\psi}|(t, \log\lambda)$.

In 2015, Irène Walspurger proved that, under a strict but feasible condition on $\boldsymbol{\psi}$, the scalogram operator is invertible up to a constant phase shift, which is inaudible [51]. Indeed, despite the loss of phase, there is enough redundancy between adjacent frequency bands λ to encode phase differences. Although Waldspurger's proof did not give a closed-form expression for $\boldsymbol{x}$ as a function of $\mathbf{U_1}(\boldsymbol{x})$, it did provide an iterative algorithm which converges to a signal $\boldsymbol{y}$ whose wavelet scalogram $\mathbf{U_1}(\boldsymbol{y})$ is equal to $\mathbf{U_1}(\boldsymbol{x})$. This surprising result harbingers a new era for digital audio effects (DAFX): developing data-driven generative models in the wavelet scalogram domain, rather than the raw waveform domain, and then sonifying the result by phase retrieval.

A second difficulty of working in the scalogram domain is that not every nonnegative function of two variables $\widetilde{\mathbf{U}}_\mathbf{1}$ is necessarily the scalogram of some real-valued waveform $\boldsymbol{x}$. This is in agreement with the Heisenberg uncertainty theorem, which prescribes a tradeoff between localization in the time domain and the Fourier domain. More precisely, the complex-valued wavelet transform underlying $\widetilde{\mathbf{U}}_\mathbf{1}$ must satisfy a reproducing kernel equation [30, Equation 4.40]. The problem of developing synthesis models in the scalogram domain while ensuring that this condition remains satisfied is, to this day, largely an open problem—but one whose solution would open new possibilities in computer music research.

4.3. *Texture synthesis*

For lack of an adequate computational framework for manipulating scalogram representations while preserving the reproducing kernel condition, one is left with two options. The first option, routinely employed in the phase vocoder and its further enhancements [40, 41, 25], is to rely on a heuristic of vertical coherence to synthesize a surrogate phase function that is, if not free of artifacts, at least perceptually plausible. The second option is to gradually morph $\boldsymbol{x}$ into a new signal whose scalogram matches the target scalogram. In particular, one may construct a trajectory of gradient descent approximations $\boldsymbol{y_0}, \ldots, \boldsymbol{y_n}$ converging towards $\boldsymbol{y}$. Once again, the legacy of Fourier is particularly insightful in this regard. Indeed, setting the initial guess $\boldsymbol{y_0}$ to match the Fourier spectrum of $\boldsymbol{x}$ produces a re-synthesis which matches the spectral envelope of the original material, but is devoid of any impulsive features. This is achieved in practice by defining $\boldsymbol{y_0}$ as

$$\boldsymbol{y_0}(t) = \mathcal{F}^{-1}\big(\omega \mapsto |\mathcal{F}(\boldsymbol{x})|(\omega) \times \exp(\mathrm{i}\boldsymbol{\varphi}(\omega)\big), \tag{11}$$

wherein the values of the phase function $\boldsymbol{\varphi}(\omega)$ are drawn at random, as independent samples from the uniform distribution in the interval $[0, 2\pi[$. Here, $\boldsymbol{y_0}$ and $\boldsymbol{x}$ have the same Fourier spectrum $|\widehat{\boldsymbol{y_0}}|(\omega) = |\widehat{\boldsymbol{x}}|(\omega)$ but differ in their wavelet spectra: $\mathbf{U_1}(\boldsymbol{y_0})(t, \log\lambda) \neq \mathbf{U_1}(\boldsymbol{x})(t, \log\lambda)$. As the iteration number n increases, the reconstruction $\boldsymbol{y}_n$ progressively exhibits transient phenomena, such as percussive onsets, loudness variations, and chirps.

For the computer musician, Fourier textures and wavelet textures stand at opposite ends of a continuum between global and local perceptions of spectral envelope. Consequently, while the former fails to recover spectrotemporal modulations, the latter is restricted to local convolutions within Q wavelet pseudo-periods or so, i.e., time scales ranging roughly between 1 ms (at $\lambda = 12\,\mathrm{kHz}$) and 100 ms (at $\lambda = 120\,\mathrm{Hz}$) for $Q = 12$. In 2010, the prospect of mitigating this conundrum led Stéphane Mallat and coauthors to develop a new family of signal representations, known as wavelet scattering transforms [31]. Although wavelet scattering was initially proposed for tasks of classification or regression, we show below that it may also find applications in signal generation, and notably musical creation.

4.4. *Joint time–frequency scattering*

The instance of wavelet scattering that we present here is known as joint time–frequency scattering. We define 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 $\boldsymbol{\psi}_\beta(t)$ even though they do not necessarily have the same shape as the wavelets $\boldsymbol{\psi}_\lambda(t)$ in Equation 9. Wavelets $\boldsymbol{\psi}_\alpha(t)$ and $\boldsymbol{\psi}_\beta(t)$ do not operate on the signal $\boldsymbol{x}$ in the time domain; rather, they perform convolutions over the temporal and frequential dimensions of the scalogram $\mathbf{U_1}(\boldsymbol{x})$. On one hand, frequencies α are measured in Hertz and discretized as 2^n with integer n. On the other hand, frequencies β are measured in cycles per octave and discretized as $\pm 2^n$ with integer n. These modulation scales β play the same role as the quefrencies in the power cepstrum.

We define the fourth-order tensor $\mathbf{U_2}(\boldsymbol{x})$ of 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:

$$\mathbf{U_2}(\boldsymbol{x})(t, \lambda, \alpha, \beta) = \big|\mathbf{U_1}(\boldsymbol{x}) \overset{t}{*} \boldsymbol{\psi}_\alpha \overset{\log_2\lambda}{*} \boldsymbol{\psi}_\beta\big|(t, \lambda) = \left| \iint_{\mathbb{R}^2} \mathbf{U_1}(\boldsymbol{x})(\tau, s)\boldsymbol{\psi}_\alpha(t-\tau)\boldsymbol{\psi}_\beta(\log_2\lambda - s)\,\mathrm{d}\tau\,\mathrm{d}s \right|, \tag{12}$$

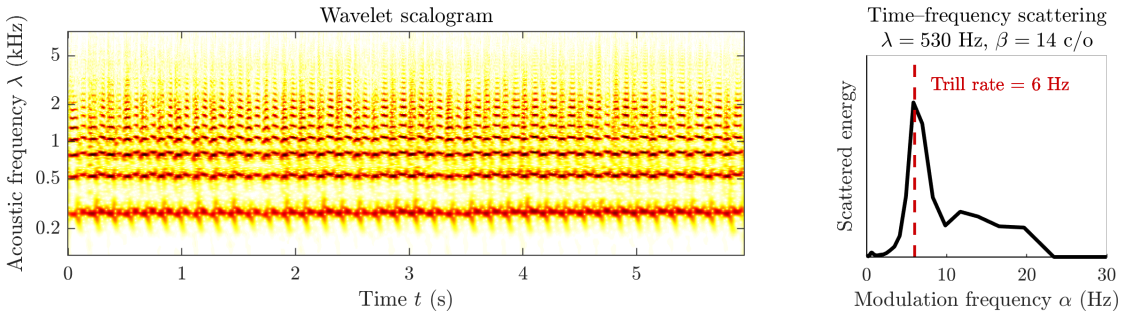

Figure 3. Left: wavelet scalogram of a trumpet playing a trill with pitch C4. Right: near the second harmonic ($\lambda = 520\,\mathrm{Hz}$, an energy peak in time–frequency scattering coefficients reveals a spectrotemporal modulation at the corresponding trill rate ($\alpha = 6\,\mathrm{Hz}$) and trill extent ($\beta = 14$ channels/octave).

Neurophysiological experiments have demonstrated that, while the wavelet scalogram $\mathbf{U_1}(\boldsymbol{x})$ can be regarded as computationally analogous to the cochlea, the tensor $\mathbf{U_2}(\boldsymbol{x})$ is a biologically plausible model for the response of the primary auditory cortex [11]. Time–frequency scattering consists in the feature-wise concatenation of $\mathbf{U_1}(\boldsymbol{x})$ and $\mathbf{U_2}(\boldsymbol{x})$ followed by local averaging over a fixed time scale T, denoted by $\mathbf{S}(\boldsymbol{x})$.

Figure 3 (left) shows the wavelet scalogram $\mathbf{U_1}(\boldsymbol{x})$ of the same trumpet trill as was presented in Figure 2. Figure 3 (right) illustrates that, in the vicinity of the second harmonic ($\lambda = 520\,\mathrm{Hz}$), and for a modulation scale β set to 14 cycles per octave (c/o), time–frequency scattering coefficients $\mathbf{S_2}(\boldsymbol{x})$ peak at a modulation frequency $\alpha = 6\,\mathrm{Hz}$ that is equal to the trill rate in $\boldsymbol{x}$. The ability of time–frequency scattering to characterize spectrotemporal modulations is discussed in greater detail in [1].

By applying a gradient descent algorithm on the Euclidean distance functional $\mathbf{E}_{\boldsymbol{x}} : \boldsymbol{y} \mapsto \|\mathbf{S}(\boldsymbol{y}) - \mathbf{S}(\boldsymbol{x})\|$, it is possible to update $\boldsymbol{y_0}$ to match the time–frequency scattering coefficients of $\boldsymbol{x}$. This algorithm was originally proposed by Joan Bruna in the simpler case of time scattering, where the operator $\mathbf{U_2}$ comprises a temporal wavelet $\boldsymbol{\psi}_\alpha$, but no frequential wavelet $\boldsymbol{\psi}_\beta$ [7]. Several theoretical results by Waldspurger, notably the invertibility of infinite-depth scattering networks [51] and the exponential decay of scattering coefficients [50], suggest that time–frequency scattering has the ability to accurately characterize longer-range dependencies T than Fourier modulus or averaged wavelet modulus coefficients. In the case of musical sounds, setting T to 50 ms or greater in the computation of spectrogram coefficients or scalogram coefficients leads to noticeable artifacts in the reconstruction. In comparison, setting T up to 500 ms yields a scattering-based reconstruction $\boldsymbol{y_n}$ that is perceptually similar to $\boldsymbol{x}$ [2].

4.5. *Composing music with wavelet scattering*

Time–frequency scattering was originally developed as a signal representation for automatic classification of audio signals. In 2016, however, composer Florian Hecker proposed to extend its application beyond the mere analysis of sounds and repurpose it towards a creative application. From an audio fragment $\boldsymbol{x}$ of duration equal to 17 s, Hecker computed a waveform $\boldsymbol{y}_0$ according to Equation 11 by matching the amplitudes between $|\widehat{\boldsymbol{x}}|$ and $|\widehat{\boldsymbol{y}}_0|$ in the Fourier domain but randomizing the corresponding phases. He then performed gradient descent to synthesize $\boldsymbol{y_1}, \ldots, \boldsymbol{y}_{50}$ iteratively, thus converging towards $\boldsymbol{x}$ in the sense of the associated Euclidean distance functional $\mathbf{E}_{\boldsymbol{x}}$, with $Q = 12$ and $T = 188\,\mathrm{ms}$. The resulting piece, named *FAVN*, was premiered at the Alte Oper in Frankfurt, presented again at the Geometry of Now festival in Moscow, and became a two-month exhibition at the Kunsthalle in Vienna, with a dedicated retrospective catalogue [28]. At the time of the concert, the performer of *FAVN* has to play back the first iteration of the first fragment, and then move forward progressively in the digital reproduction of the piece, both in terms of compositional time (fragments) and computational time (iterations). Since then, Hecker has composed three original pieces with time–frequency scattering: *Modulator (Scattering Transform)* (2012), *Experimental Palimpsest* (2016), and *Inspection* (2016).

Beyond the aesthetic of experimental music, time–frequency scattering has recently found a wider audience by appearing on an electronica record named *The Shape of RemiXXXes to Come*, released by the independent

record label Warp [29]. In his remix, Hecker isolated a few one-bar loops from Lorenzo Senni's *XAllegroX* and reconstructing them from their time–frequency scattering coefficients. While the Fourier-based initial guess $\boldsymbol{y}_0$ sounds hazy and static, the reconstruction regains some of the original rhythmic content in subsequent iterations, thus producing a sensation of sonic rise. In the context of dance music, such a sensation conveys the anticipation of a sudden changepoint, or "drop". To produce a rise, one widespread audio engineering technique consists in isolating a percussive sample and playing it repeatedly while reducing progressively the duration of this sample. A second technique consists in applying a high-pass filter whose cutoff frequency increases progressively through time. These two techniques operate in the time domain and in the Fourier domain respectively. In contrast, time–frequency scattering achieve a sensation of rise by affecting the spectrotemporal organization of sound, from coarse to fine, independently from pure time–domain or frequency–domain manipulations. The release of *XAllegroX (scattering.m remix)* subverts the classical articulation of tension and release in dance music by leaving both the perception of musical meter and the perception of musical register unchanged, and, instead, modulating the spectrotemporal complexity of texture itself.

5. Conclusion

The democratization of digital audio storage in the 1980s, followed by the massive adoption of Internet communication in the 1990s, intensified the need for efficient computational models of auditory perception. From compressive coding to speech enhancement, many of the technologies enabling wireless audio networking rely on the fast Fourier transform or FFT—hailed by the journal *IEEE Computing in Science & Engineering* as one of the ten most important algorithms of the 20^{th} century. Thus, the ubiquity of digital signal processing algorithms in industrialized societies is proof that the scientific legacy of Fourier remains, on the 250^{th} anniversary of his birth, highly relevant for addressing the mathematical challenges of the information age.

Astonishingly the publication of the *Analytical Theory of Heat* (1822) by Joseph Fourier precedes the invention of digital computers by over a hundred years. Despite this timespan, the theory of trigonometric series described therein is foundational in the history of digital signal processing. The case of musical acoustics exemplifies the longevity of Fourier representations in many applied domains of science. Indeed, the introduction of Fourier series constitutes a watershed in our understanding of timbre—i.e., how the shape and playing technique of an instrument influences our qualitative perception of the sound it produces, aside from pitch and intensity. Practically speaking, the main appeal behind Fourier decompositions lies in the imitation of acoustic instruments with a small, physically interpretable set of continuous parameters.

In his treatise, Fourier pointed out that his proposed method for solving the heat equation on a metal rod could also be applied, *mutatis mutandis*, to solve the d'Alembert equation on a vibrating string. While the initial application of this framework was merely analytical, the advent of computer music in the 1960s introduced the practical use of Fourier series in both phases of sound modeling: analysis (i.e., visualizing and classifying pre-recorded sounds) and synthesis (i.e., transforming pre-recorded sounds and generating new sounds). Moreover, the coordinated efforts of researchers and composers have demonstrated that Fourier series can also generate previously unheard timbres and thus serve as an ergonomic interface for contemporary music creation.

In music production, an extensive array of digital audio effects (DAFX) relies on some form of spectrotemporal analysis and re-synthesis after manipulation in the time–frequency domain. One well-known example of this procedure is the phase vocoder, at the heart of the "Auto-Tune" algorithm for vocal pitch adjustment. The availability of these DAFX have not only inspired composers with new ideas, but has also created new technological needs. In particular, although the current state of the art for Fourier-based analysis–synthesis now faithfully replicates the ordinary instrumentarium, it is inadequate for replicating sounds that are considered neither speech nor music, e.g., domestic sounds or wildlife vocalizations. More generally, the question of modeling audio textures in the absence of any prior knowledge of temporal periodicity remains largely open. There is a need for multiresolution analysis tools that can hardly be implemented using a single-resolution Fourier analysis scheme. Far from purely academic, this question represents a technical bottleneck on the way towards even more intuitive computer–human interactions.

In the last section of this article, we have presented time–frequency scattering, a nonlinear representation

of audio signals which, alongside other recent models based on spectrotemporal modulations, offers new possibilities for texture synthesis. Following the methodological tradition of early computer music research, in which art and science serve complementary roles, the development of time–frequency scattering was associated with the creation of new pieces, both in so-called "avant-garde" and "pop" aesthetics. Time–frequency scattering is heavily inspired by Fourier analysis because it relies on convolutions with Morlet wavelets, which are well-localized both in the time domain and in the Fourier domain. However, time–frequency scattering also borrows from other paradigms of more recent inception, and notably deep convolutional networks. Future research in this direction will investigate the relationship between time–frequency analysis and nonlinear multiresolution analysis in two contexts: nonlinear multiresolution analysis for the processing of natural sounds, and the synthesis of relevant material for the music of our time.

Acknowledgment

This work is supported by the ERC InvariantClass grant 320959. The Flatiron Institute is a division of the Simons Foundation.

References

[1] J. Andén, V. Lostanlen, and S. Mallat. "Joint Time–Frequency Scattering". In: *IEEE Trans. Sig. Process.* 67.14 (2019), pp. 3704–3718.

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

[3] M. Andreatta, F. Nicolas, and C. Alunni. *À la lumière des mathématiques et à l'ombre de la philosophie. Actes du séminaire mamuphi*. Ircam/Delatour France, 2012.

[4] H. Berjamin et al. "Time-domain numerical modeling of brass instruments including nonlinear wave propagation, viscothermal losses, and lips vibration". In: *Acta. Acust. united Ac.* 103.1 (2017), pp. 117–131.

[5] M. Betser et al. "Estimation of frequency for AM/FM models using the phase vocoder framework". In: *IEEE Transactions on Signal Processing* 56.2 (2008), pp. 505–517.

[6] B. Bogert, M. Healy, and J. Tukey. "The quefrency alanysis of time series for echoes. Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking". In: *Proceedings of a Symposium on Time Series Analysis*. John Wiley & Sons, Inc., 1963, pp. 209–243.

[7] J. Bruna and S. Mallat. "Audio texture synthesis with scattering moments". In: *arXiv preprint arXiv:1311.0407* (2013).

[8] M. Castellengo. *Écoute musicale et acoustique*. Eyrolles, 2015.

[9] J. M. Chowning. "The synthesis of complex audio spectra by means of frequency modulation". In: *J. Audio Eng. Soc.* 21.7 (1973), pp. 526–534.

[10] A. Cont. "ANTESCOFO: Anticipatory Synchronization and Control of Interactive Parameters in Computer Music". In: *Proc. ICMC*. 2008, pp. 33–40.

[11] D. Depireux et al. "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex". In: *J. Neurophysiol.* 85.3 (2001), pp. 1220–1234.

[12] H. Dudley. "The vocoder—Electrical re-creation of speech". In: *J. Soc. Motion Pict. Eng.* 34.3 (1940), pp. 272–278.

[13] K. Fitz, L. Haken, and P. Christensen. "A new algorithm for bandwidth association in bandwidth-enhanced additive sound modeling". In: *Proc. ICMC*. 2000, pp. 384–387.

[14] J. L. Flanagan et al. "Phase vocoder". In: *J. Acoust. Soc. Am.* 38.5 (1965), pp. 939–940.

[15] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer-Verlag, 1991.

[16] J. Fourier. *Théorie analytique de la chaleur*. Bibliothèque nationale de France, no. FRBNF30454516, 1822.

[17] R. M. Friedman. "The creation of a new science: Joseph Fourier's analytical theory of heat". In: *Hist. Stud. Phys. Sci.* 8 (1977), pp. 73–99.

[18] J. Harvey. *The Music of Stockhausen: An Introduction*. University of California Press, 1975.
[19] A. Herreman. "L'inauguration des séries trigonométriques dans la *Théorie analytique de la chaleur de Fourier* et dans la controverse des cordes vibrantes". In: 19.2 (2013), pp. 151–243.
[20] F. Jedrzejewski. *Mathématiques des sysèmes acoustiques*. L'Harmattan, 2002.
[21] M. Kac. "Can one hear the shape of a drum?" In: *Am. Math. Mon.,* 73.4P2 (1966), pp. 1–23.
[22] W. Koenig, H. Dunn, and L. Lacy. "The sound spectrograph". In: *J. Acoust. Soc. Am.* 18.1 (1946), pp. 19–49.
[23] R. Kronland-Martinet, J. Morlet, and A. Grossmann. "Analysis of sound patterns through wavelet transforms". In: *Int. J. Pattern Recognit. Artif. Intell.* 1.02 (1987), pp. 273–302.
[24] J. Laroche and M. Dolson. "Improved Phase Vocoder Time–scale Modification of Audio". In: *IEEE Trans. Speech. Audio Process.* 7.3 (1999), pp. 323–332.
[25] J. Laroche and M. Dolson. "New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing and other exotic audio modifications". In: *J Audio Eng. Soc.* 47.11 (1999), pp. 928–936.
[26] W.-H. Liao, A. Roebel, and A. W.-Y. Su. "On the Modeling of Sound Textures Based on the STFT Representation". In: *Proc. DAFx*. 2013.
[27] M. Liuni et al. "Automatic Adaptation of the Time–Frequency Resolution for Sound Analysis and Re-Synthesis". In: *IEEE Trans. Audio Lang. Speech Process.* 21.5 (2013), pp. 959–970.
[28] V. Lostanlen. "On time–frequency scattering and computer music". In: *Florian Hecker: Halluzination, Perspektive, Synthese*. Ed. by V. J. M. Nicolaus Schafhausen. Berlin: Sternberg Press, 2019.
[29] V. Lostanlen and F. Hecker. "*The Shape of RemiXXXes to Come*: Audio texture synthesis with time-frequency scattering". In: *Proc. DAFx*. 2019.
[30] S. Mallat. *A wavelet tour of signal processing: The sparse way*. Academic Press, 2009.
[31] S. Mallat. "Understanding deep convolutional networks". In: *Phil. Trans. R. Soc. A* 374.2065 (2016), p. 20150203.
[32] M. V. Mathews. "The Digital Computer as a Musical Instrument". In: *Science* 142.3592 (1963), pp. 553–557.
[33] M. V. Mathews et al. *The technology of computer music*. Vol. 969. MIT press Cambridge, 1969.
[34] S. McAdams et al. "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes". In: *Psychol. Res.* 58.3 (1995), pp. 177–192.
[35] J. H. McDermott and E. P. Simoncelli. "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis". In: *Neuron* 71.5 (2011), pp. 926–940.
[36] M. Mersenne. *Harmonie universelle, contenant la théorie et la pratique de la musique*. Bibliothèque nationale de France, no. FRBNF30932210, 1636.
[37] E. S. Ottosen et al. "A phase vocoder based on nonstationary Gabor frames". In: *IEEE Trans. Audio Speech Lang. Process.* 25.11 (2017), pp. 2199–2208.
[38] J.-C. Risset. *Du songe au son. Entretiens avec Matthieu Guillot*. L'Harmattan, 2008.
[39] J.-C. Risset. "Synthèse de sons à l'aide de calculateurs électroniques appliquée à l'étude de sons de trompette". In: *Comptes Rendus de l'Académie des Sciences de Paris* B.263 (1966), pp. 111–114.
[40] A. Röbel. "A new approach to transient processing in the phase vocoder". In: *Proc. DAFx*. 2003.
[41] A. Röbel. "A shape-invariant phase vocoder for speech transformation". In: *Proc. DAFx*. 2010.
[42] X. Rodet and P. Depalle. "Spectral envelopes and inverse FFT synthesis". In: *Proc. AES*. 1992.
[43] P. Schaeffer. *In search of a concrete music*. Vol. 15. University of California Press, 2012.
[44] P. Schaeffer. *Treatise on Musical Objects: An Essay across Disciplines*. University of California Press, 2017.
[45] M. R. Schroeder. "Auditory paradox based on fractal waveform". In: *J. Acoust. Soc. Am.* 79.1 (1986), pp. 186–189.
[46] X. Serra and J. O. Smith. "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition". In: *Comp. Mus. J.* 14.4 (1990), pp. 12–24.
[47] C. Shannon. "Communication in the presence of noise". In: *Proc. Inst. Radio Eng.* 37.1 (1949), pp. 10–21.
[48] R. Strachan. *Sonic Technologies: Popular Music, Digital Culture and the Creative Process*. Bloomsbury Publishing USA, 2017.

[49] H. Triebel. *Theory of Function Spaces*. Birkhaüser Verlag, 1992.
[50] I. Waldspurger. "Exponential decay of scattering coefficients". In: *Proc. SampTA*. IEEE. 2017, pp. 143–146.
[51] I. Waldspurger. "Wavelet transform modulus: phase retrieval and scattering". PhD thesis. Ecole normale supérieure, 2015.
[52] T. Wishart. "The Composition of *Vox-5*". In: *Comput. Music J.* 12.4 (1988), pp. 21–27.
[53] U. Zölzer. *DAFX: Digital audio effects*. John Wiley & Sons, 2011.