THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'École normale supérieure

Convolutional Operators in the Time-frequency Domain
Opérateurs convolutionnels dans le plan temps-fréquence

École doctorale n°386

SCIENCES MATHÉMATIQUES DE PARIS CENTRE

Spécialité INFORMATIQUE

COMPOSITION DU JURY :

M. GLOTIN Hervé
LSIS, AMU, Université de Toulon, ENSAM, CNRS, président du jury

M. PEETERS Geoffroy
STMS, Ircam, Université Pierre et Marie Curie, CNRS, rapporteur

M. RICHARD Gaël
LTCI, TELECOM ParisTech, Université Paris-Saclay, CNRS, rapporteur

M. MALLAT Stéphane
DI, École normale supérieure, CNRS, membre du jury

M. LAGRANGE Mathieu
IRCCyN, École centrale de Nantes, CNRS, membre du jury

M. SHAMMA Shihab
LSP, École normale supérieure, CNRS, membre du jury

Soutenue par VINCENT **LOSTANLEN**
le 2 février 2017

Dirigée par **Stéphane MALLAT**

ENS
PSL RESEARCH UNIVERSITY PARIS

CONVOLUTIONAL OPERATORS IN THE TIME-FREQUENCY DOMAIN

VINCENT LOSTANLEN

Département d'informatique
École normale supérieure

Vincent Lostanlen: *Convolutional operators in the time-frequency domain*.
This work is supported by the ERC InvariantClass grant 320959.

In memoriam Jean-Claude Risset, 1938-2016.

ABSTRACT

In the realm of machine listening, audio classification is the problem of automatically retrieving the source of a sound according to a predefined taxonomy. This dissertation addresses audio classification by designing signal representations which satisfy appropriate invariants while preserving inter-class variability. First, we study time-frequency scattering, a representation which extracts modulations at various scales and rates in a similar way to idealized models of spectrotemporal receptive fields in auditory neuroscience. We report state-of-the-art results in the classification of urban and environmental sounds, thus outperforming short-term audio descriptors and deep convolutional networks. Secondly, we introduce spiral scattering, a representation which combines wavelet convolutions along time, along log-frequency, and across octaves, thus following the geometry of the Shepard pitch spiral which makes one full turn at every octave. We study voiced sounds as a nonstationary source-filter model where both the source and the filter are transposed in frequency through time, and show that spiral scattering disentangles and linearizes these transpositions. In practice, spiral scattering reaches state-of-the-art results in musical instrument classification of solo recordings. Aside from audio classification, time-frequency scattering and spiral scattering can be used as summary statistics for audio texture synthesis. We find that, unlike the previously existing temporal scattering transform, time-frequency scattering is able to capture the coherence of spectrotemporal patterns, such as those arising in bioacoustics or speech, up to a scale of about 500 ms. Based on this analysis-synthesis framework, an artistic collaboration with composer Florian Hecker has led to the creation of five computer music pieces.

PUBLICATIONS

1. Lostanlen, V and S Mallat (2015). "Transformée en scattering sur la spirale temps-chroma-octave". In: *Actes du GRETSI.*

2. Andén, J, V Lostanlen, and S Mallat (2015). "Joint Time-frequency Scattering for Audio Classification". In: *Proceedings of the IEEE Conference on Machine Learning for Signal Processing (MLSP)*. Received a Best Paper award.

3. Lostanlen, V and S Mallat (2015). "Wavelet Scattering on the Pitch Spiral". In: *Proceedings of the International Conference on Digital Audio Effects (DAF-x)*.

4. Lostanlen, V and C Cella (2016). "Deep Convolutional Networks on the Shepard Pitch Spiral for Musical Instrument Recognition". In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)*.

5. Lostanlen, V, G Lafay, J Andén, and M Lagrange (2017). "Auditory Scene Similarity Retrieval and Classification with Relevance-based Quantization of Scattering Features". To appear in *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), Special Issue on Sound Scene and Event Analysis*.

ACKNOWLEDGMENTS

First and foremost, I thank Stéphane Mallat, my advisor, for being a day-to-day supporter of the "try it nonlinear and diagonal" method (Mallat, 2008, preface), which goes well beyond the realm of applied mathematics.

For accepting to be part of my PhD committee, I thank Hervé Glotin, Mathieu Lagrange, Geoffroy Peeters, Gaël Richard, and Shihab Shamma.

For fostering my interest in science from an early age, I thank Christophe Billy, Thomas Chénel, Bernard Gauthier, Sébastien Guffroy, Raphaëla López, Étienne Mahé, Joaquin Martinez, and Anna Poveda.

For teaching me the fundamentals of audio signal processing, I thank Roland Badeau, Bertrand David, Slim Essid, and Jérémie Jakubowicz, who were my professors at TELECOM ParisTech.

For supervising me as an intern in information geometry at Inria, and providing me a pleasant foretaste of scientific research, I thank Arshia Cont and Arnaud Dessein.

For introducing me to the theory of the time-frequency scattering transform, and helping me putting it into practice, I thank Joakim Andén, who, more than a co-author, has been my mentor throughout these years of doctoral studies. I also thank Joan Bruna, who was the first in applying the scattering transform to audio texture synthesis, and was kind enough to explain his approach to me.

I thank Carmine-Emanuele Cella who co-authored an ISMIR paper with me on deep convolutional networks on the pitch spiral; and Grégoire Lafay, who contributed with me to a TASLP paper on relevance-based quantization of scattering features.

I thank composer Florian Hecker for his relentless enthusiasm in using my software for computer music applications. I also thank Bob Sturm for putting us in contact.

For their cordial invitations to give talks, I thank Lucian Alecu, Moreno Andreatta, Elaine Chew, Anne-Fleur Multon (from TrENSmissions student radio station), Ophélie Rouby, Simon Thomas (from Union PSL association), and Bruno Torrésani.

For the small talk and the long-winded conversations, I thank my colleagues from École normale supérieure: Mathieu Andreux, Tomás Angles, Mickaël Arbel, Mia Xu Chen, Xiuyan Cheng, Ivan Dokmanic, Michael Eickenberg, Sira Ferradans, Matthew Hirn, Michel Kapoko, Bangalore Ravi Kiran, Chris Miller, Edouard Oyallon, Laurent Sifre, Paul Simon, Gilles Wainrib, Irène Waldspurger, Guy Wolf, Zhuoran Yang, Tomás Yany, and Sixin Zhang.

I thank the PhD students from other labs who interacted with me: Thomas Andrillon, Pablo Arias, Victor Bisot, Stanislas Chambon, Keunwoo Choi, Léopold Crestel, Philippe Cuvillier, Pierre Donat-Bouillud, Simon Durand, Edgar Hemery, Clément Laroche, Simon Leglaive, Juan Ulloa, and Neil Zeghidour.

I was invited to review the work of four interns: Randall Balestriero, Alice Cohen-Hadria, Christian El Hajj, and Florian Menguy. I wish them fair winds and following seas in their scientific careers.

For their alacrity in solving computer-related problems, even when the problem lies between the chair and the keyboard, I thank Jacques Beigbeder and the rest of the SPI (which stands for "Service des Prodiges Informatiques") at École normale supérieure.

At the time of this writing, I am a postdoctoral researcher affiliated to the Cornell Lab of Ornithology, working at the Center for Urban Science and Progress (CUSP) as well as the Music and Audio Research Lab (MARL) of New York University. I thank Juan Pablo Bello, Andrew Farnsworth, and Justin Salamon, for hiring me into this exciting position. More broadly, I thank Rachel Bittner, Mark Cartwright, Andrea Genovese, Eric Humphrey, Peter Li, Jong Wook Kim, Brian McFee, Charlie Mydlarz, and Claire Pelofi, for their welcome.

I thank Iain Davies, Marshall Iliff, Heather Wolf, and the rest of the Information Department at the Cornell Lab for introducing me to the world of birdwatching and wildlife conservation.

Finally, I thank my family and friends for their unfailing support. My most heartfelft thoughts go to my parents, Isabelle and Michel Lostanlen, and to my girlfriend, Claire Vernade.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

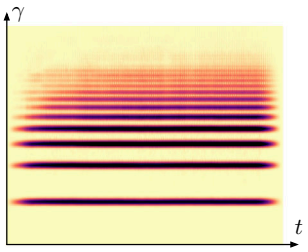
1 INTRODUCTION

This PhD dissertation addresses the problem of audio classification, that is, of identifying sources from a single-channel recording according to a predefined taxonomy. Audio classification is one of the earliest applications of machine listening, the branch of computer science dedicated to constructing an artificial intelligence of sounds (Rowe, 1992).

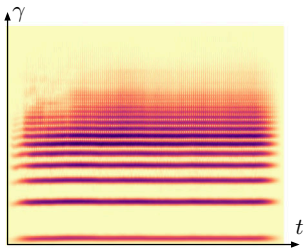
Evaluating an audio classification system is only possible if the mutual agreement between human raters is sufficiently high to produce a dataset of annotated signals, called ground truth. Such a ground truth is not available for all perceptual attributes of music information (Schedl, Gómez, and Urbano, 2014), whose appraisal inevitably entails a part of subjectivity. Yet, unlike genres or chords, musical instruments can be associated to sound files among a digital music corpus in an unequivocal way. This is because the taxonomy of instruments proceeds naturally from the fundamental principles of musical acoustics (Fletcher and Rossing, 2012). Likewise, the taxonomy of environmental sounds is derived from the various kinds of mechanical interactions emitting sound (Salamon, Jacoby, and Bello, 2014). In this dissertation, we focus on improving the performance of automated systems for the classification of environmental sounds (chapter 3) and musical instruments (chapter 4).

The challenge of musical instrument classification is made difficult by the large intra-class variability induced by expressive performance. Be them hand-crafted or learned from data, signal representations strive to reduce this intra-class variability while preserving enough inter-class variability to discriminate classes. Because, at fine temporal scales, most of the intra-class variability can be modeled as a local deformation in time and in frequency, such signal representations are most often defined in the time-frequency domain, that is, after a short-term Fourier transform (STFT) spectrogram or a continuous wavelet transform (CWT) scalogram (Flandrin, 1998).

Figure 1.1a shows the wavelet scalogram of a musical note as played by a trumpet, indexed by time t and log-frequency γ. The next subfigures in Figure 1.1 show the scalograms of the same human-instrument interaction, under the effect of many independent factors of variability: pitch, intensity, attack, tone quality, tonguing, sordina, articulation, and phrasing. It appears that each of these factors affect the spectrotemporal evolution of the signal at scales ranging from the duration of an onset (20 milliseconds) to the duration of the full musical note (2 seconds).

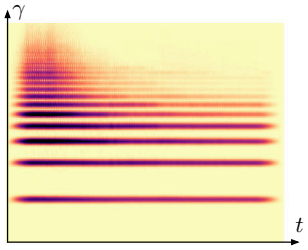


(a) Original note (G_4).

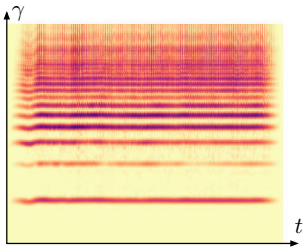


(b) Variation in pitch (G_3).

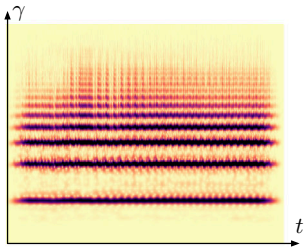
(c) Variation in intensity (*pianissimo*).



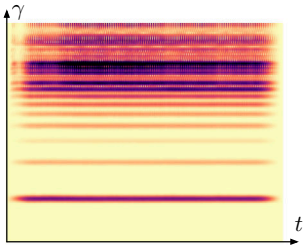
(d) Variation in attack (*sforzando*).



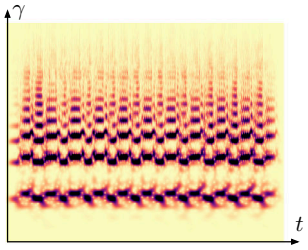
(e) Variation in tone quality (*bright*).



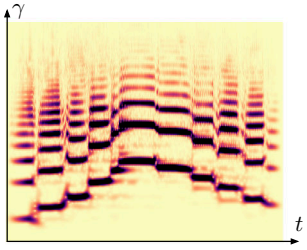
(f) Variation in tonguing (*flatterzunge*).



(g) Variation in sordina (*harmon*).



(h) Variation in articulation (*trill*).



(i) Variation in phrasing (*arpeggio*).

Figure 1.1: Nine variations in music. Subfigure (a) shows the scalogram of a trumpet note. Subfigures (b) to (i) show some variations of (a) along independent parameters: pitch, intensity, attack, tone quality, tonguing, sordina, articulation, and phrasing. All sounds proceed from the Studio OnLine (SOL) dataset.

A desirable representation for audio classification should meet two requirements. First, it must *segregate* factors of variability into linearly independent dimensions (Bengio, 2013). Secondly, it must *integrate* the fast modulations of auditory information into a slower evolution (Atlas and Shamma, 2003). In this dissertation, we present a new signal representation, named time-frequency scattering (TFS), which aims at combining both of these properties.

Time-frequency scattering associated with a linear support vector machine classifier (SVM) yields state-of-the-art results in urban sound classification and musical instrument recognition, thus outperforming currently existing approaches based on short-term audio descriptors as well as deep convolutional networks (ConvNets). Moreover, we show that audio waveforms can be re-synthesized from the feature space and that they bear a strong perceptual resemblance with the original signal.

Before going into further details, we describe the remaining challenges for audio classification in the next section. Section 1.2 lists the main contributions and Section 1.3 provides an outline of the dissertation.

1.1 CURRENT CHALLENGES

The main challenge of audio classification is to disentangle the factors of variability in every signal from the training set. Such factors of variability may be geometrical (e.g. time shifts), acoustical (e.g. frequency transpositions), or musical (e.g. expressivity). Once this is achieved, a supervised classifier can build an invariant representation by ruling out factors of intra-class variability while retaining factors of inter-class variability.

In this section, we bring three different perspectives to this problem. The first one relates to temporal integration of short-term descriptors, the second one relates to the characterization of non-Gaussian stationary processes, and the third one relates to harmonic correlations along the frequency axis.

1.1.1 *Beyond short-term audio descriptors*

The timbre of a musical instrument is essentially determined by its shape and its materials. Such properties remain constant through time. Therefore, musical instruments, like many other human-controlled objects producing sound, can be modeled as dynamical systems with time-invariant parameters.

Yet, the dependency between the amplitude of the input excitation and the amplitude of the output signal is often nonlinear. As a result, sharp onsets produce distinctive time-frequency patterns, whose temporal structure is not taken into account by short-term audio de-

scriptors (20 milliseconds), such as spectral centroid or spectral flux. Integrating the transientness of these patterns up to the time scale of a musical phrase (2 seconds) without losing their temporal structure would improve the generalization power of spectrotemporal invariant representations.

In this dissertation, we address the challenge of temporal integration by cascading two levels of wavelet modulus operators. The former yields a well-known time-frequency representation, called scalogram, which segregates acoustic frequencies from low to high. The latter yields a three-way tensor, indexed by time, log-frequency, and a third variable corresponding to temporal or spectrotemporal modulations. Temporal scattering results from temporal modulations, while time-frequency scattering results from spectrotemporal modulations.

For a broad class of natural sounds, averaging this tensor through time yields an invariant representation which is sparser than the averaged scalogram. In other words, extracting modulations from the scalogram segregates factors of variability into linearly independent dimensions, hence a gain in inter-class discriminability at long temporal scales (Chi, Ru, and Shamma, 2005).

1.1.2 *Beyond weak-sense stationarity*

The re-synthesis of an audio signal from its invariant representation can be interpreted as the statistical sampling of a stationary process from empirical measurements of some of its ergodic quantities. Measuring the average short-term Fourier spectrum, interpreted as a power spectral density, merely recovers first-order and second-order moments of the stationary process. This method is sufficient for "static" sounds (e.g. humming, buzzing) but fails to re-synthesize longer-range interactions in auditory textures (e.g. roaring, crackling). Striving to improve the realism of the reconstruction while increasing the time scale of local stationarity in the target signal provides a test bed for the comparison of invariant audio descriptors which is free of any bias in dataset, task, or classifier. However, this test bed ultimately rests upon the judgment of human listeners, who may disagree in electing the closest match to the target. Therefore, the comparative evaluation of invariant representations by means of signal re-synthesis does not necessarily lead to a clear result as to which representation is the best.

The state of the art in audio texture synthesis is currently held by McDermott and Simoncelli (2011), who extracted summary statistics from an invariant representation akin to the temporal scattering transform. These summary statistics encompass the empirical mean, variance, skewness, and kurtosis of every coefficient, as well as local cross-correlations in frequency. A gradient descent algorithm re-synthesizes a new signal with similar summary statistics as the orig-

inal texture. This method is sufficient for sounds with independent amplitude dynamics (e.g. falling rain, applause) at the time scale of 20 ms but fails to re-synthesize more intricate time-frequency structures (e.g. tweeting, talking) beyond the time scale of 200 ms. Furthermore, because it relies on high-order statistical moments, this invariant representation is unstable to intra-class variability, and would thus be inadequate for classification.

In this dissertation, we re-synthesize sounds from local averages of temporal scattering and time-frequency scattering. We find that the reconstruction obtained from temporal scattering, although more "dynamic" than the average Fourier spectrum, falls behind the method of McDermott and Simoncelli (2011). Indeed, it misaligns percussive onsets spanning across adjacent wavelet subbands. Yet, time-frequency scattering, which extracts empirical means of spectrotemporal modulations instead of merely temporal modulations, mitigates this issue. Therefore, time-frequency scattering is on par with the state of the art in audio texture synthesis, without having to account for high-order moments.

1.1.3 *Beyond local correlations in frequency*

The previous decade has witnessed a breakthrough of deep learning, and in particular deep convolutional networks (ConvNets), in audio signal processing. A deep convolutional network is a stack of learned convolutional kernels interspersed with pointwise nonlinearities and local pooling operators. As depth increases, convolutional kernels reach larger spectrotemporal scales and their behavior tends to gain in abstraction.

Yet, the assumption that all informative correlations are concentrated to local time-frequency patches hinders the discriminative power of ConvNets for audio signals, which contain harmonic combs in the Fourier domain. A harmonic comb spans across the whole log-frequency axis from its fundamental frequency to its topmost partials, with a decreasing distance between consecutive partials.

It stems from the above that a dataset of pitched spectra is not stationary along the log-frequency axis. Moreover, empirical correlations are not limited to local neighborhoods in frequency, but also appear at common musical intervals, such as octaves, perfect fifths, and major thirds. Exploring the geometry of these correlations, and providing well-adapted solutions, could help raising the "glass ceiling" encountered by feature engineering (Aucouturier and Pachet, 2004) as well as the current state of the art in feature learning (Choi, Fazekas, and Sandler, 2016).

In this thesis, we roll up the log-frequency axis into a spiral which makes one turn at every octave, such that power-of-two harmonics get aligned on a same radius. As a result, empirical correlations

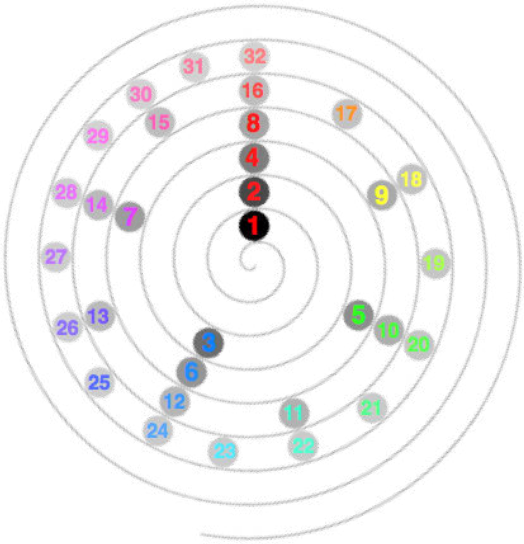


Figure 1.2: Once the log-frequency axis has been rolled up into a spiral which makes one turn at every octave, power-of-two harmonics get aligned on the same radius. Pitch chroma and pitch height are respectively denoted by hue and brightness.

are concentrated on a three-dimensional neighborhood of time, log-frequencies, and octaves. We show that performing convolutions across octaves improves the classification accuracy of deep convolutional networks as well as time-frequency scattering.

1.2 CONTRIBUTIONS

The main contribution of this dissertation is to define new convolutional operators in the time-frequency domain with applications in audio classification and texture synthesis. These operators are built by applying a cascade of wavelet transforms over multiple variables onto a time-frequency representation and applying complex modulus. The three variables investigated in this dissertation are time, log-frequency, and discrete octave index.

This dissertation proceeds incrementally in the definition of convolutional operators. First, temporal scattering, which was previously introduced by Andén and Mallat (2014), is described in Section 3.1. Secondly, time-frequency scattering, which is known in auditory neuroscience as spectrotemporal receptive fields (STRF) (Chi, Ru, and Shamma, 2005), is described in Section 3.2. Thirdly, spiral scattering, which is a thoroughly novel contribution, is described in Section 4.2. The former two of these operators are suited to all kinds of audio sig-

nals, whereas spiral scattering is mostly relevant in the case of pitched sounds, in which harmonic intervals are especially salient.

Along with the definition of each of the aforementioned convolutional operators, we intend to provide an insight on its capabilities and limitations in terms of representational power. To this aim, we combine three methodologies: theoretical analysis of invariants; qualitative appraisal of re-synthesized textures; and quantitative evaluation in a supervised classification setting.

1.2.1 *Theoretical analysis of invariants*

First, we express the variability of natural sounds as a class of smooth, local deformations in the time-frequency domain, and establish how the response of the convolutional operator is affected by such deformations. All scattering operators satisfy invariance to time shifts (Subsection 2.2.1) and stability to small time warps (Subsection 3.1.3). Yet, only time-frequency scattering and spiral scattering are sensitive to time reversal (Subsection 2.2.4) and frequency-dependent time shifts (Subsections 3.1.4 and 3.2.3).

Then, we focus on the case of pitched sounds by studying the scattering coefficients of the stationary source-filter model (Subsection 2.4.1) as well as some of its nonstationary generalizations. All scattering operators are able to retrieve coherent amplitude modulations (Subsection 3.1.3), but only spiral scattering is able to disentangle and linearize the nonstationary contributions of source and filter, thus segregating the effects of pitch and timbral modifications through time (Subsections 4.3.2 and 4.3.3).

1.2.2 *Qualitative appraisal of re-synthesized textures*

Secondly, we develop an algorithm to synthesize audio signals whose time-averaged scattering coefficients match the summary statistics of a natural audio texture (Subsection 3.3.2). Then, we compare perceptually the re-synthesized signal with the original for each kind of scattering representation, and derive an appraisal on its ability to segregate and integrate auditory information.

We find that time-frequency scattering, unlike temporal scattering, accurately synchronizes frequency subbands (Subsection 3.3.3). Moreover, spiral scattering brings a slight improvement in the recovery of harmonic structures and broadband impulses with respect to time-frequency scattering (Subsection 4.4.2). Besides its primary purpose as a tool for the comparison of scattering representations, our algorithm is employed in the field of computer music to generate new sounds (Subsection 3.3.4).

1.2.3 *Supervised classification*

Thirdly, we evaluate the discriminative power of scattering coefficients in several tasks of audio classification. The state of the art in this domain is held by spectrogram-based feature learning techniques, and deep convolutional networks (ConvNets) in particular (Humphrey, Bello, and Le Cun, 2013). Because they result from a cascade of two convolutional operators interspersed with pointwise modulus nonlinearities, scattering networks have the same architecture as a two-layer convolutional network. However, the convolutional kernels in a scattering network are set as wavelets instead of being learned from data.

We consider three kinds of audio classification datasets: short environmental sounds (Piczak, 2015b; Salamon, Jacoby, and Bello, 2014), 30-second acoustic scenes (Stowell et al., 2015), and continuous recordings of musical instruments (Bittner et al., 2014; Joder, Essid, and Richard, 2009). In all of them, temporal scattering is outperformed by deep convolutional networks with spectrotemporal kernels. Yet, replacing temporal scattering by time-frequency scattering achieves state-of-the-art results in environmental sound classification (Subsection 3.4.4) and musical instrument recognition (Subsection 4.5.4). Replacing time-frequency scattering by spiral scattering provides a slight improvement in the classification accuracy of musical instruments.

1.3 OUTLINE

The rest of this dissertation is organized as follows.

1.3.1 *Time-frequency analysis of signal transformations*

In Chapter 2, we address the problem of revealing signal transformations of increasing abstraction by means of time-frequency analysis. We begin with *geometrical* transformations which have a closed-form expression in the time domain, such as shifts, dilations, nonlinear warps, and time reversal. Then, we describe *acoustical* transformations, such as frequency transposition, which are ill-defined in the time domain but may be defined as affine transformations in the time-frequency domain. Lastly, we review the *musical* transformations of some abstract parameters of sound, such as pitch, timbre, and rhythm.

1.3.2 *Time-frequency scattering*

In Chapter 3, we address the problem of improving the accuracy of the currently existing approaches for the description of auditory textures, such as urban sounds and acoustic scenes. We define temporal

scattering and discuss its connections with previous research on amplitude modulation features. Then, we define time-frequency scattering, discuss its connections with previous research on spectrotemporal modulation features, and compare its discriminative ability with respect to temporal scattering. We present a fast implementation of time-frequency scattering, released as open source software. We apply automatic differentiation to derive a tractable gradient descent algorithm for any kind of scattering transform, be it temporal or spectrotemporal. We use gradient descent to perform audio texture synthesis and present a creative application with composer Florian Hecker. We report state-of-the-art performance in urban sound classification and competitive performance in acoustic scene classification.

1.3.3 *Spiral scattering*

In Chapter 4, we address the problem of musical instrument recognition in solo phrases. We present theoretical arguments, corroborated by music theory and auditory neuroscience, in favor of rolling up the log-frequency axis into a spiral which makes a full turn at every octave. We define spiral scattering, a refinement of time-frequency scattering, extracting modulations along time, log-frequency, and across octaves. We define a nonstationary generalization of the source-filter model, in which amplitude, pitch, and brightness jointly vary through time, and show how spiral scattering disentangles these factors of variability. We report state-of-the-art performance in musical instrument classification, thus outperforming deep convolutional networks.

2

TIME-FREQUENCY ANALYSIS OF SIGNAL TRANSFORMATIONS

Time-frequency analysis is a preliminary step in the design of invariant representations for audio classification. In this chapter, we review the fundamental tools of time-frequency analysis with a focus on the wavelet modulus operator. Assuming that the signal locally follows asymptotic conditions of pseudo-periodicity, we give an approximate equation for its wavelet transform and discuss the implications of perceptual transformations in the time-frequency domain.

Also known as scalogram, the wavelet modulus operator is a two-dimensional function of time and log-frequency. Section 2.1 defines the scalogram along with other time-frequency operators for audio classification, such as the short-term Fourier transform (STFT) and the nonstationary Gabor transform (NSGT).

Small affine transformations of an audio signal do not affect its class. Because wavelets are obtained by shifts and dilations, these transformations in the signal domain remain affine in the scalogram. Section 2.2 explains how the wavelet modulus operator demodulates pseudo-periodic oscillations, thus constructing an invariant to small affine transformations.

Frequency transposition, a transformation that often does not affect the class of interest, is intuitively understood as as motion from low to high frequencies and vice versa. However, because of Fourier duality, it cannot be defined purely in the frequency domain without undesirable effects in the time domain. Section 2.3 is devoted to frequency transposition, time stretching, and variation in spectral envelope, three notions that are ill-defined in the time domain but defined as affine transformations in the time-frequency domain.

Musicians refer to sound according to a variety of "parameters" such as pitch, timbre, and rhythm. These notions have acoustical correlates, but it is admittedly hazardous, if ever possible, to explicit them with a closed-form model. Section 2.4 emphasizes the difficulties of grasping high-level musical transformations from the scalogram, while highlighting what is at stake for audio classification.

Part of the content of this chapter has been previously published in (Lostanlen and Cella, 2016).

2.1 TIME-FREQUENCY REPRESENTATIONS

Musical notation exposes the pitch and duration of every note in a piece. Whereas durations can only be measured in the time domain,

pitches can only be measured in the frequency domain. The problem of instrument classification is to build an invariant representation to pitches and durations while remaining sensitive to timbre, that is, to the physical principles underlying musical acoustics.

Disentangling pitches from durations, as a preliminary step to the construction of an invariant representation for classification, is the purpose of time-frequency analysis. This section reviews the two most common time-frequency representations, namely the short-term Fourier transform (STFT, Subsection 2.1.1) and the continuous wavelet transform (CWT, Subsection 2.1.2).

Both of these two classical tools are less than ideal for audio signal processing: the STFT lacks temporal localization at high frequencies and frequential localization at low frequencies, and vice versa for the CWT. To find a compromise between the STFT and the CWT, we describe the nonstationary Gabor transform (NSGT) of Balazs et al. (2011) in 2.1.3, which will be used as time-frequency representation in the following chapters.

2.1.1 *Short-term Fourier transform*

In this subsection, we introduce the Fourier transform operator in $L^2(\mathbb{R}, \mathbb{R})$. We show that the modulus of the Fourier transform is invariant to translations, but lacks redundancy to discriminate classes. We define Gabor time-frequency atoms $\boldsymbol{\psi}_{\omega,T}(t)$ of center frequency ω and time scale T. Lastly, we define the short-term Fourier transform and the Fourier spectrogram.

Fourier transform

Within a framework of continuous time, sound is encoded by a pressure wave $\boldsymbol{x} \in L^2(\mathbb{R}, \mathbb{R})$ of finite energy

$$\|x\|_2 = \sqrt{\int |\boldsymbol{x}|^2(t)\, \mathrm{d}t}. \tag{2.1}$$

Sinusoidal waves, which are at the foundation of Fourier analysis, are the eigenvectors of convolutional operators (Mallat, 2008, Theorem 2.2). The Fourier transform operator

$$\hat{\boldsymbol{x}}(\omega) = \int_{-\infty}^{+\infty} \boldsymbol{x}(t) \exp(-2\pi \mathrm{i}\omega t)\, \mathrm{d}t \tag{2.2}$$

is defined in the space $L^1(\mathbb{R}, \mathbb{R})$ of integrable functions, and then extended into $L^2(\mathbb{R}, \mathbb{R})$ by density (Mallat, 2008, Section 2.2). Its functional inverse is

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{\boldsymbol{x}}(\omega) \exp(2\pi \mathrm{i}\omega t)\, \mathrm{d}\omega. \tag{2.3}$$

Translation operator

For some $b \in \mathbb{R}$, let

$$\mathcal{T}_b : \boldsymbol{x}(t) \longmapsto (\mathcal{T}_b \boldsymbol{x})(t) = \boldsymbol{x}(t+b) \tag{2.4}$$

be the translation operator. Applying the Fourier transform to the translated signal $(\mathcal{T}_b \boldsymbol{x})$ yields

$$\widehat{\mathcal{T}_b \boldsymbol{x}}(\omega) = \exp(-2\pi \mathrm{i} \omega b) \times \hat{\boldsymbol{x}}(\omega), \tag{2.5}$$

that is, the Fourier transform $\hat{\boldsymbol{x}}(\omega)$ up to a phase term $\exp(-2\pi \mathrm{i} \omega b)$. The Fourier spectrum of $\boldsymbol{x}(t)$ is defined as the magnitude coefficients $|\hat{\boldsymbol{x}}|(\omega)$ of $\hat{\boldsymbol{x}}(\omega)$. It follows from the equation above that all translations $(\mathcal{T}_b \boldsymbol{x})(t)$ of $\boldsymbol{x}(t)$ have the same Fourier spectrum. This is a desirable property, because $\boldsymbol{x}(t)$ and $(\mathcal{T}_b \boldsymbol{x})(\omega)$ belong to the same class. Nevertheless, because the Fourier transform lacks redundancy, the reciprocal is not true: two signals with similar Fourier spectra may not be translated from one another. The inadequacy of the Fourier spectrum for audio classification is particularly noticeable at large temporal scales, i.e. typically above 20 ms.

Gabor time-frequency atoms

To circumvent this issue, audio descriptors are defined over short-term windows and subsequently aggregated through time by means of machine learning techniques. Let

$$\boldsymbol{g}_T(t) = \frac{1}{T} \exp\left(-\frac{t^2}{2T^2}\right)$$

be a Gaussian window function of typical support ranging between $t = -3T$ and $t = +3T$. We denote by

$$\boldsymbol{\psi}_{\omega,T}(t) = \boldsymbol{g}_T(t) \exp(2\pi \mathrm{i} \omega t).$$

the Gabor time-frequency atom of center frequency ω and time scale T.

Fourier spectrogram

For every region (t, ω) of the time-frequency plane, the short-term Fourier transform (STFT) is defined as

$$\begin{aligned} \mathrm{STFT}(\boldsymbol{x})(t,\omega) &= (\boldsymbol{x} * \boldsymbol{\psi}_{\omega,T})(t) \\ &= \int_{-\infty}^{+\infty} \boldsymbol{x}(t') \boldsymbol{g}_T(t-t') \exp\left(2\pi \mathrm{i} \omega (t-t')\right) \, \mathrm{d}t' \end{aligned}$$

The modulus of the short-term Fourier transform, called spectrogram, remains almost unchanged by the action of the translation $\mathcal{T}_b$ as long

as b remains small in front of T. Consequently, the spectrogram is locally invariant to translation, but globally covariant.

The choice of time scale T results from an empirical tradeoff between temporal localization and spectral localization. For most applications, the audio signal processing community has settled on approximately $T = 50\,\text{ms}$. On one hand, this value enables to discriminate percussive onsets as long as they are distant of at least T, that is, if no more than $\frac{1}{2T} = 10$ onsets per second are present in the analyzed signal. On the other hand, the frequential localization is of the order of 10 Hz, that is, approximately a semitone (one twelfth of an octave) around middle C (261 Hz).

2.1.2 *Continuous wavelet transform*

In this subsection, we argue in favor of multiresolution analysis of audio signals instead of adopting a single time scale T. We introduce Gabor wavelets $\boldsymbol{\psi}_{\omega,\frac{Q}{\omega}}(t)$ of center frequency ω, time scale $\frac{Q}{\omega}$, and constant quality factor Q. We define the Gabor scalogram as the complex modulus of the continuous wavelet transform. We discretize frequencies ω according to a geometric sequence 2^γ, where the log-frequency variable γ takes evenly spaced values.

Multiresolution analysis of musical transients

A musical note typically consists of two parts: a transient state and a steady state. The transient state, also called attack, has a sharp temporal localization and a coarse frequential localization. Conversely, the steady state, which encompasses sustain and release, has a precise fundamental frequency but no precise offset time. The short-term Fourier transform adopts a single time scale T to analyze the transient state and the steady state. As such, it fails to describe the temporal evolution of musical attacks, which have transient structures at scales finer than T. Yet, musical attacks have proven to be of crucial importance in psychophysical experiments of instrument recognition (Grey and Gordon, 1978).

Wavelets

The problem of characterizing musical attacks as well as pitched spectra cannot be solved with Gabor atoms of constant scales T. However, designing a family of linear time-frequency atoms of constant shape

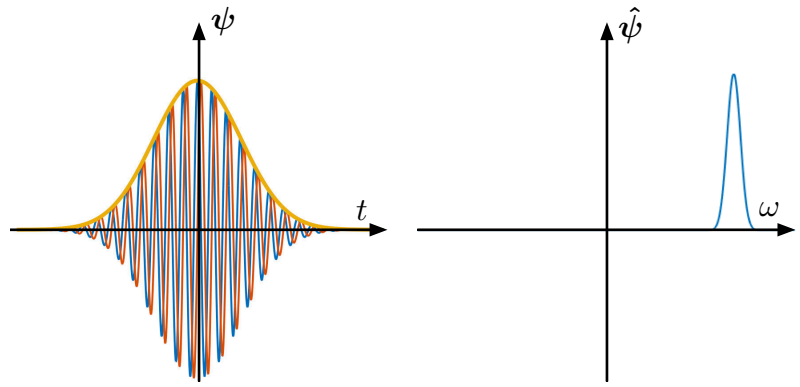
$$\boldsymbol{\psi}_{\omega,\frac{Q}{\omega}}(t) = \boldsymbol{g}_{\frac{Q}{\omega}}(t)\exp(2\pi\mathrm{i}\omega t)$$

yields a time-frequency representation in which the time scale grows in inverse proportion with the center frequency. Such time-frequency atoms of constant shape are named wavelets or constant-Q filters. We refer to Daubechies (1996) for a short historical perspective on

wavelets, and to the textbook of Mallat (2008) for an in-depth introduction.

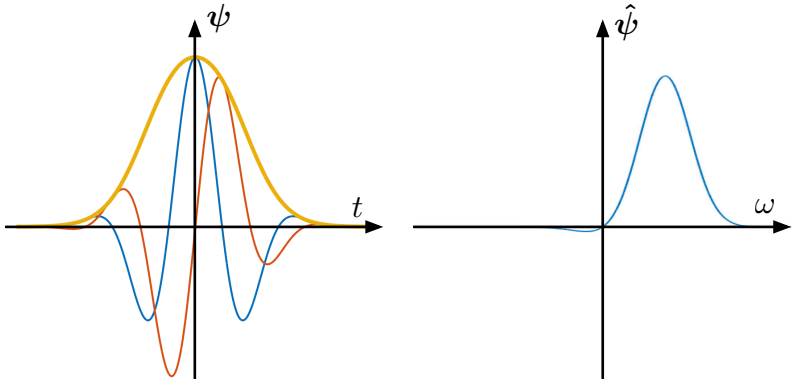
The wavelet representation is not subject to the choice of a time scale T, but of a quality factor Q, that is, a Heisenberg tradeoff in time-frequency localization. The quality factor of $\psi_{\omega,\frac{Q}{\omega}}(t)$ is of the same order of magnitude as its number of non-negligible oscillations, which is independent from the choice of ω.

Two Gabor atoms with different quality factors are shown in Figure 2.1, both in the time domain and in the Fourier domain.



(a) Gabor wavelet, $Q = 8$, time domain.

(b) Gabor wavelet, $Q = 8$, Fourier domain.



(c) Gabor wavelet, $Q = 1$, time domain.

(d) Gabor wavelet, $Q = 1$, Fourier domain.

Figure 2.1: Gabor wavelets with different quality factors: (a) $Q = 8$, time domain ; (b) $Q = 8$, Fourier domain ; (c) $Q = 1$, time domain ; (d) $Q = 1$, Fourier domain. Blue and red oscillations respectively denote real and imaginary part, while the orange envelope denotes complex modulus.

Wavelet scalogram

The continuous wavelet transform (CWT), also called constant-Q transform (CQT), of $\boldsymbol{x}(t)$ is defined as

$$\begin{aligned} \mathrm{CWT}(t,\omega) &= (\boldsymbol{x} * \boldsymbol{\psi}_{\omega,\frac{Q}{\omega}})(t) \\ &= \int_{-\infty}^{+\infty} \boldsymbol{x}(t')\boldsymbol{g}_{\frac{Q}{\omega}}(t-t')\exp(2\pi\mathrm{i}\omega(t-t'))\,\mathrm{d}t' \end{aligned}$$

The modulus of the continuous wavelet transform is called wavelet scalogram.

Geometric sequence of center frequencies

The interval between two pitches is characterized by the ratio of their fundamental frequencies. Consequently, musical intervals are defined along a perceptual scale which is approximately logarithmic in frequency. In numerical applications and displays, center frequencies ω of the wavelet transform are discretized according to a geometric sequence

$$\omega = 2^{\gamma},$$

where the log-frequency variable

$$\gamma = \log_2 \omega$$

takes evenly spaced values. Wavelets have a temporal support of $2^{\gamma}Q$, a center frequency of $2^{-\gamma}$, a bandwidth of $2^{-\gamma}/Q$, and a constant quality factor of Q.

The choice of integer 2 as the base for the geometric sequence 2^{γ} facilitates numerical implementations, which are based on the Fast Fourier Transform (FFT) (Beylkin, Coifman, and Rokhlin, 1991; Cooley and Tukey, 1965; Van Loan, 1992). Since a musical octave corresponds to a dilation factor of 2, incrementing γ to $(\gamma+1)$ amounts to an upward interval of one octave, so the integer part of γ can be interpreted as an octave index. Setting the quality factor Q to 12, or a multiple thereof, matches the Western tradition of twelve-tone equal temperament (Jedrzejewski, 2002).

2.1.3 *Nonstationary Gabor transform*

In this subsection, we address the problem of striking a compromise between the time-frequency localizations of the short-term Fourier transform (STFT) and the continuous wavelet transform (CWT). We withdraw into classical experiments in psychoacoustics to settle whether a combination of two pure tones will be segregated in frequency or integrated through time. We give an empirical formula for the equivalent rectangular bandwidth (ERB) of the human ear as a function of

the acoustic frequency. We introduce the nonstationary Gabor transform (NSGT), a generalization of the STFT and the CWT. We visualize all three representations (STFT, CWT, NSGT) as time-frequency images, and verify that the NSGT provides sharper edges throughout the whole hearing range.

Auditory segregation of neighboring tones

Let ξ_A and ξ_B be two frequencies in the typical hearing range 100 Hz - 10 kHz. Let

$$\begin{aligned} x(t) &= x_A(t) + x_B(t) \\ &= \cos(2\pi\xi_A t) + \cos(2\pi\xi_B t) \end{aligned} \tag{2.6}$$

be the superposition of two sine waves of respective frequencies ξ_A and ξ_B. A trigonometric formula allows to rewrite the signal x as

$$\begin{aligned} x(t) &= 2\cos\left(2\pi\frac{\xi_B+\xi_A}{2}t\right)\cos\left(2\pi\frac{\xi_B-\xi_A}{2}t\right) \\ &= 2\cos(2\pi\xi_1 t)\cos(2\pi\xi_2 t) \end{aligned} \tag{2.7}$$

that is, an amplitude modulation (AM) signal of carrier frequency $\xi_1 = \frac{\xi_B+\xi_A}{2}$ and modulation frequency $\xi_2 = \frac{\xi_B-\xi_A}{2}$.

Equations 2.6 and 2.7 are mathematically interchangeable. However, in the realm of computational auditory scene analysis (CASA), they do not account for the same phenomenon. In Equation 2.6, the pure tones $x_A(t)$ and $x_B(t)$ are segregated in frequency, whereas in Equation 2.7, they are integrated through time as a "beating" wave.

Equivalent rectangular bandwidths

Whether the auditory system performs segregation or grouping depends upon the values of ξ_A and ξ_B. In the mammalian auditory system, the basilar membrane of the cochlea can be modeled as a set of auditory filters which performs spectrotemporal segregation of acoustic events. The bandwidths of these filters can be measured experimentally by exposing human subjects to ambiguous signals $x = x_A(t) + x_B(t)$ and asking whether the pure tones $x_A(t)$ and $x_B(t)$ interfere or not.

The equivalent rectangular bandwidth (ERB) at ξ_A is defined as the range of values in ξ_B, called critical band, such that $x(t)$ is heard as one beating tone instead of two steady-state tones. Because it is conditional upon experimental validation, it varies across authors. A widespread rule of thumb (Glasberg and Moore, 1990) is the affine function

$$\mathrm{ERB}(\omega) = 24.7 + 0.108 \times \omega.$$

Necciari et al. (2013) have used the function $\mathrm{ERB}(\omega)$ to define an "ERBlet transform" in which the time scale of the Gabor atom at the center frequency ω is proportional to the function

$$T(\omega) = \frac{1}{\mathrm{ERB}(\omega)} = \frac{1}{24.7 + 0.108 \times \omega},$$

that is, the inverse of the equivalent rectangular bandwidth.

General framework

In full generality, the collection of Gabor atoms

$$\boldsymbol{\psi}_{\omega,T(\omega)}(t) = \boldsymbol{g}_{T(\omega)}(t) \exp(2\pi \mathrm{i} \omega t)$$

with center frequencies ω and time scales $T(\omega)$ defines a nonstationary Gabor transform

$$\mathrm{NSGT}(\boldsymbol{x})(t,\omega) = (\boldsymbol{x} * \boldsymbol{\psi}_{\omega,T(\omega)})(t)$$

where the time scales $T(\omega)$ are adapted to the center frequencies ω. We refer to Balazs et al. (2011) for an introduction to the theory of the nonstationary Gabor transform (NSGT), and in particular how to ensure energy conservation as well as invertibility.

The pointwise complex modulus

$$\mathbf{U}_1\boldsymbol{x}(t,\gamma) = |\boldsymbol{x} * \boldsymbol{\psi}_{2^\gamma,T(2^\gamma)}|(t)$$

of the NSGT, called scalogram, remains almost unchanged by the action of the translation $\mathcal{T}_b$ as long as b remains small in front of $T(\omega)$. In other words, the scalogram is rather invariant (resp. covariant) to translation at lower (resp. higher) acoustic frequencies ω.

Mel scale

Our choice for the function $T(\omega)$ is

$$T(\omega) = \frac{1}{\max\left(\frac{\omega}{Q_{\max}}, \frac{1}{T_{\max}}\right)}.$$

The corresponding Gabor transform is a short-term Fourier transform below the cutoff frequency $\frac{Q_{\max}}{T_{\max}}$ and a continuous wavelet transform above that frequency. This choice is inspired by the mel scale (Umesh, Cohen, and Nelson, 1999), designed by psychoacoustical experiments such that perceptually similar pitch intervals appear equal in width over the full hearing range (Stevens, Volkmann, and Newman, 1937).

Visualizations of scalograms as time-frequency images

Figure 2.2 shows the audio waveform of a jazz recording, along with three time-frequency representations: the STFT, the CWT, and the NSGT. To facilitate the comparison between visualizations, the vertical axis of the STFT is a log-frequency axis $\gamma = \log_2 \omega$ instead of a frequency axis. Near the lowest frequencies, we observe that the STFT lacks frequential discriminability whereas the CWT lacks temporal discriminability. The NSGT strikes a good compromise between the two representations.

In all following applications and figures, the auditory transform we use is a NSGT of maximal quality factor $Q_{\max} = 24$ and maximal window length $T_{\max} = 100\,\text{ms}$.

2.2 GEOMETRICAL TRANSFORMATIONS

A single-channel audio signal is a real-valued function $x(t)$ of the time variable t. A geometrical transformation of the signal is defined by a change of variable

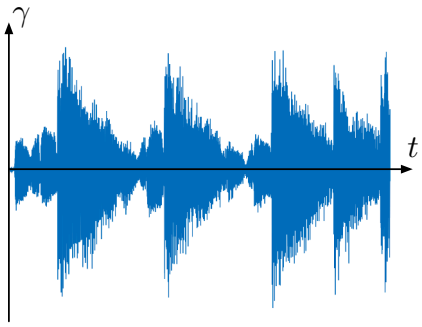
$$(\mathcal{W}_\tau x)(t) = \dot{\tau}(t) \times (x \circ \tau)(t),$$

where $\tau(t)$ is a time warp function and $\dot{\tau}(t)$ denotes the temporal derivative of $\tau(t)$. Whether the auditory patterns in $x(t)$ are affected by the time warp $\tau(t)$ depends on the temporal context in $x(t)$ and of the local regularity of $\tau(t)$.

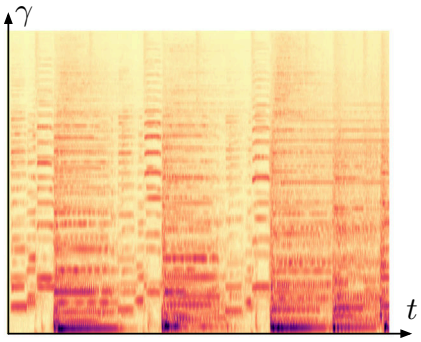
In this section, we address the problem of modeling geometrical transformations of audio signals by means of time-frequency analysis. First of all, we restrict ourselves to the case of affine transformations $\tau(t) = at + b$, which form the foundation of the group-theoretical insights behind wavelet theory. Subsection 2.2.1 is devoted to time shifts $\tau(t) = t + b$ whereas Subsection 2.2.2 is devoted to time dilations $\tau(t) = at$. Subsection 2.2.3 leaves the special case of affine transformations to address all kinds of slowly varying, nonlinear time warps. Recalling a theoretical result of Delprat et al. (1992), we show that the analytic amplitude of a locally monochromatic signal can be retrieved from its wavelet scalogram. Lastly, Subsection 2.2.4 introduces Gammatone wavelets, which, unlike Gabor wavelets, are sensitive to time reversal.

2.2.1 *Time shifts*

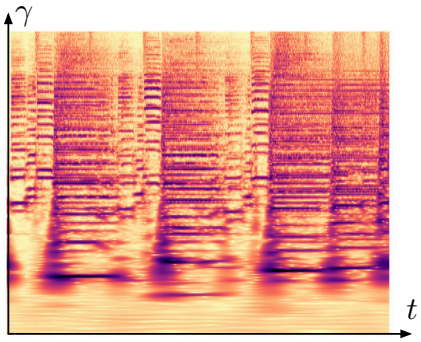
In this subsection, we study the particular case $\tau(t) = t + b$ where $b \in \mathbb{R}$, that is, the geometrical transformation $\mathcal{W}_\tau$ boils down to the time shift $\mathcal{T}_b$. In the context of music information retrieval, we provide a formal distinction between translation-covariant tasks (detection) and translation-invariant tasks (classification). Then, we relativize this



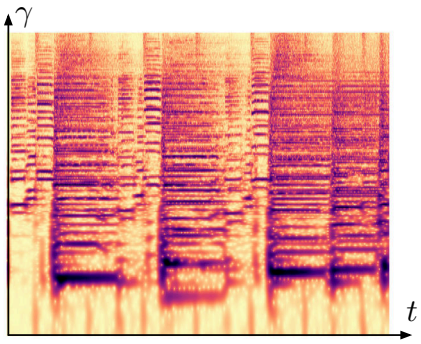
(a) Audio waveform.



(b) Short-term Fourier transform (STFT).



(c) Continuous wavelet transform (CWT).



(d) Nonstationary Gabor transform (NSGT).

Figure 2.2: Three time-frequency representations, indexed by time t and log-frequency $\gamma = \log_2 \omega$.

distinction by pointing out that the hierarchical levels of music information form a nested hierarchy of time scales. We define the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ and the averaged scalogram $\mathbf{S}_1\boldsymbol{x}(t,\gamma)$, the latter being invariant to translations of at most T. We relate the notions of regularity along time and sparsity across features, which are paramount to efficient classification, to grouping and segregation in computational auditory scene analysis. We give a probabilistic interpretation of the limit case $T \to \infty$, corresponding to a global invariance to time shifts.

Translation-covariant vs. translation-invariant tasks

Time shifts affect the locations of auditory patterns without changing their assigned classes. The vast majority of tasks in audio-based information retrieval dissociate location from content, and can thus be gathered in two categories: translation-covariant tasks, i.e. detection ; and translation-invariant tasks, i.e. classification.

For example, translation-covariant tasks include sound onset detection (Bello et al., 2005), beat tracking (Ellis, 2007), and chord estimation (McVicar et al., 2014). In contrast, translation-invariant tasks include musical genre recognition (Sturm, 2014b), acoustic scene classification (Stowell et al., 2015), musical instrument recognition (Joder, Essid, and Richard, 2009), and tag retrieval (Eck et al., 2007).

In the former, the expected outcome of the system should contain the temporal cues associated to each event of interest. In the latter, only a global description of the information in $\boldsymbol{x}(t)$ is required. Applying $\mathcal{T}_b$ to a signal $\boldsymbol{x}(t)$ results in the translated waveform $\boldsymbol{x}(t+b)$. It subtracts b to the temporal cues in detection tasks and leaves the labels unchanged in classification tasks.

The focus on translation is also motivated by the fact that the temporal organization of audio signals is generally not known. For instance, in acoustic scene classification, the starting time $t = 0$ of $\boldsymbol{x}(t)$ is chosen arbitrarily by the recordist, and thus does not convey any information about the class. The same applies to speech and music as long as there is no time-aligned symbolic transcription available, that is, in the most challenging settings.

Hierarchy of time scales in auditory information

One should bear in mind that the notion of invariance to translation is always subordinated to the choice of a time scale T. Indeed, invariant information at a small scale may become covariant at a larger scale. The case of music, in which hierarchical levels of information are superimposed, is compelling in that regard (Vinet, 2003). Elementary stimuli, such as percussive onsets, can be resolved at time scales as small as 20 milliseconds (Polfreman, 2013). Onset retrieval is thus formulated as multi-object detection, and invariance to translation is restricted to $T = 50$ ms. Nevertheless, considering instrument activa-

tions ($T = 500$ ms), genres ($T = 5$ s), and musical forms ($T > 50$ s) as nested levels of information, it appears that a gradual increase in musical abstraction is accompanied by a gradual increase in the orders of magnitude of T. Depending on the duration of the query and the level of abstraction to be achieved, the corresponding task is either cast as detection or classification.

Averaged scalogram

Recalling Subsection 2.1.3, we define the scalogram of an audio waveform $\boldsymbol{x}(t)$ as the matrix

$$\mathbf{U}_1\boldsymbol{x}(t,\gamma) = |\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^\gamma, T(2^\gamma)}|(t),$$

indexed by time t and log-frequency γ. A translation-invariant representation $\mathbf{S}_1\boldsymbol{x}(t,\gamma)$ up to time shifts $b \ll T$ can be linearly obtained from $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ by convolving it with a low-pass filter $\boldsymbol{\phi}_T(t)$ of cut-off frequency set to $1/T$, thus performing a moving average over the variable t for every wavelet subband γ of center frequency 2^γ:

$$\mathbf{S}_1\boldsymbol{x}(t,\gamma) = (\mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\phi}_T)(t,\gamma) = \int_{-\infty}^{+\infty} \mathbf{U}_1\boldsymbol{x}(t',\gamma)\boldsymbol{\phi}_T(t-t')\,\mathrm{d}t'. \quad (2.8)$$

It results from the above that, assuming b lower than T, the Euclidean distance between $\mathbf{S}_1\boldsymbol{x}(t,\gamma)$ and $\mathbf{S}_1\mathcal{T}_b\boldsymbol{x}(t,\gamma)$ is small in front of the energy of the original signal, as measured by its L^2 norm:

$$\|\mathbf{S}_1\boldsymbol{x}(t,\gamma) - \mathbf{S}_1\mathcal{T}_b\boldsymbol{x}(t,\gamma)\|_2 \ll \|\mathbf{S}_1\boldsymbol{x}(t,\gamma)\|_2\,. \quad (2.9)$$

Regularity along time and sparsity across features

Temporal integration is highly beneficial for machine learning applications, because it entails a linear dimensionality reduction, hence a reduction of statistical variance and an improved generalization power. Consequently, it is desirable to make the time scale T as large as possible, i.e. up to the duration of auditory objects in the task at hand. The inevitable downside of low-pass filtering is that transient information in the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ at finer scales than T are lost, hence a lack of discriminability in feature space.

This remark does not only apply to scalograms, but also to any kind of multidimensional time series $\mathbf{U}_1\boldsymbol{x}(t,\lambda_1)$ indexed by time t and an unstructured feature index λ_1. In the sequel, $\mathbf{U}_1\boldsymbol{x}$ is the scalogram of $\boldsymbol{x}$, so that the literals λ_1 and γ both denote log-frequency, and can be used interchangeably.

If the representation $\mathbf{U}_1\boldsymbol{x}(t,\lambda_1)$ is regular along time t and sparse across features λ_1, the information in $\mathbf{U}_1\boldsymbol{x}(t,\lambda_1)$ is retained in $\mathbf{S}_1\boldsymbol{x}(t,\lambda_1)$, with the additional guarantee of invariance to translation. Regularity in time is a problem of temporal integration, whereas sparsity across features is a problem of segregation. In the next chapter, we

will show how temporal scattering and time-frequency scattering improve on both of these aspects with respects to the averaged scalogram $\mathbf{S}_1 \boldsymbol{x}(t, \lambda_1)$.

Global invariance to time shifts

When T approaches infinity, the local averaging with $\boldsymbol{\phi}_T(t)$ is no longer necessary, as it is replaced by a full delocalization

$$\mathbf{S}_1 \boldsymbol{x}(\lambda_1) = \int_{-\infty}^{+\infty} \mathbf{U}_1 \boldsymbol{x}(t', \lambda_1) \, \mathrm{d}t'. \tag{2.10}$$

The limit $T \to +\infty$ is epitomized by the task of acoustic scene classification, where the problem amounts to retrieving in which location the recording was made. Owing to the scarcity of salient events in many natural soundscapes, acoustic scene classification is only possible by integrating a large temporal context. Ideally, the translation-covariant representation $\mathbf{U}_1 \boldsymbol{x}(t, \lambda_1)$ should map distinct acoustic sources to distinct features λ_1, so that $\mathbf{S}_1 \boldsymbol{x}(\lambda_1)$ summarizes the relative presence of each source. In this case, $\mathbf{S}_1 \boldsymbol{x}(\lambda_1)$ is not only invariant to a global time shifts of the audio signal $\boldsymbol{x}(t)$, but also to relative time shifts of one acoustic source with respect to one another. This remark will be reiterated in Subsection 3.1.4 as a limitation of the temporal scattering transform.

Within a probabilistic framework, the integral above can be interpreted as the empirical estimation of the expected value of ergodic measurements $\mathbf{U}_1 \boldsymbol{X}(t, \lambda)$ for some stationary process $\boldsymbol{X}(t)$ whose observed realizations are denoted by $\boldsymbol{x}(t)$:

$$\mathbf{S}_1 \boldsymbol{X}(\lambda) = \mathbb{E}\left[\mathbf{U}_1 \boldsymbol{X}(t, \lambda)\right].$$

This perspective is particularly useful in texture classification as well as texture synthesis — see Section 3.3.

2.2.2 *Time dilations*

In this subsection, we model geometrical deformations $\mathcal{W}_\tau$ such that the time warp $\tau(t) = at + b$ is an affine function. We define the special affine group G of shifts and dilations over the real line. We draw a strong connection between G and the construction of a continuous wavelet transform.

Linear time warps

A diffeomorphism τ of the form $\tau(t) = at$ where $a > 0$ induces a dilation of the time axis by a factor a, i.e. a linear time warp. If the dilated signal $\mathcal{W}_\tau \boldsymbol{x}(t) = \boldsymbol{x}(at)$ were to be re-annotated by a human expert, its annotation at $t = 0$ would be the identical to the annotation of $\boldsymbol{x}(t)$ at $t = 0$ as long as $|1 - a| \ll 1$. Conversely, bringing a

further apart from 1 could yield a different annotation or an unrealistic sound that would be impossible to annotate. Modeling dilations is essential to audio fingerprinting, as they account for tape speed variability in analog recordings (Fenet, Richard, and Grenier, 2011). In addition, dilations provide a rudimentary way to model the frequency transpositions of "static" sounds, i.e. weak-sense stationary processes.

Special affine group

Composing shifts with dilations yields the special affine group

$$G = \left\{ \boldsymbol{\tau}_{(a,b)} : t \mapsto at + b \mid a \in \mathbb{R}_+^*, b \in \mathbb{R} \right\} \tag{2.11}$$

of direct, affine transformations of the real line. Observe that G is closed under functional composition, and that its neutral element is the identity $\boldsymbol{\tau}_{(1,0)} : t \mapsto t$. Alleviating notations, we denote by

$$\mathcal{W}_{(a,b)}\boldsymbol{x}(t) = \mathcal{W}_{\boldsymbol{\tau}_{(a,b)}}\boldsymbol{x}(t) = a \times \boldsymbol{x}(at + b) \tag{2.12}$$

the signal resulting from the deformation of $\boldsymbol{x}(t)$ by $\boldsymbol{\tau}_{(a,b)} \in G$.

Wavelets and affine transformations

Let us now consider the wavelet $\boldsymbol{\psi}(t) \in L^2(\mathbb{R}, \mathbb{C})$ of dimensionless center frequency $\omega = 1$. By an affine change of variable, it appears that filtering the dilated signal $\mathcal{W}_{(a,b)}\boldsymbol{x}(t)$ is equivalent to dilating the filtered signal $(\boldsymbol{x} * \mathcal{W}_{(\frac{1}{a},0)}\boldsymbol{\psi})(t)$, where $\mathcal{W}_{(\frac{1}{a},0)}\boldsymbol{\psi}(t) = \frac{1}{a}\boldsymbol{\psi}(\frac{t}{a})$ corresponds to a dilation of $\boldsymbol{\psi}(t)$ by a scaling factor $\frac{1}{a}$:

$$\begin{aligned}
(\mathcal{W}_{(a,b)}\boldsymbol{x} * \boldsymbol{\psi})(t) &= \int_{-\infty}^{+\infty} \boldsymbol{x}(t')\boldsymbol{\psi}\left(t - \frac{t'-b}{a}\right) \mathrm{d}t' \\
&= a \times \left(\boldsymbol{x} * \mathcal{W}_{(\frac{1}{a},0)}\boldsymbol{\psi}\right)(at + b) \\
&= \mathcal{W}_{(a,b)}\left(\boldsymbol{x} * \mathcal{W}_{(\frac{1}{a},0)}\boldsymbol{\psi}\right)(t).
\end{aligned}$$

The equation above shows that affine transformations of the time axis can be transferred from the signal $\boldsymbol{x}(t)$ to the wavelet $\boldsymbol{\psi}(t)$. As a result, convolving $\boldsymbol{x}(t)$ with all wavelets of the form $\mathcal{W}_{(\frac{1}{a},0)}\boldsymbol{\psi}(t)$ gives an insight on the group orbit

$$G\boldsymbol{x} = \left\{ \mathcal{W}_{(a,b)}\boldsymbol{x} : t \longmapsto a \times \boldsymbol{x}(at + b) \;\middle|\; a \in \mathbb{R}_+^*, b \in \mathbb{R} \right\}, \tag{2.13}$$

which is an infinite set of translated and deformed versions of $\boldsymbol{x}(t)$, as localized in time and frequency by $\boldsymbol{\psi}(t)$. The wavelet $\boldsymbol{\psi}(t)$ plays an analogous role to a measurement device in physics (Auger et al., 2013).

2.2.3 *Time warps*

In this subsection, we address the general case of geometrical transformations $\mathcal{W}_\tau$ in the time domain.

Nonlinear time warps

Besides affine transformations of the time axis $\boldsymbol{\tau}_{(a,b)} : t \mapsto at + b$, nonlinear time warps are of particular interest for audio signal processing. They are found in nature under the form of Doppler effects, but most importantly, they can model chirps (Flandrin, 2001), i.e. smooth variations in the fundamental frequencies of quasi-periodic signals, as well as variations of spectral brightness in unvoiced, weak-sense stationary sounds.

Hilbert transform and asymptotic signal

Let $\boldsymbol{x}(t) = \cos(2\pi t)$ be a sine wave of dimensionless frequency equal to 1. Warping $\boldsymbol{x}(t)$ by $\boldsymbol{\tau}(t)$ yields the chirp

$$\mathcal{W}_\tau \boldsymbol{x}(t) = \dot{\boldsymbol{\tau}}(t) \cos\left(2\pi \boldsymbol{\tau}(t)\right)$$

of instantaneous frequency $\dot{\boldsymbol{\tau}}(t)$. A classical way to estimate $\dot{\boldsymbol{\tau}}(t)$ is to demodulate the oscillations in $\mathcal{W}_\tau \boldsymbol{x}(t)$ by taking its analytic part and applying complex modulus. The analytic part of a real signal $\boldsymbol{x}(t)$ is defined as

$$\boldsymbol{x}_\mathbf{a}(t) = \boldsymbol{x}(t) + \mathrm{i} \times (\mathcal{H}\boldsymbol{x})(t) \tag{2.14}$$

where $\mathcal{H}$ denotes the Hilbert transform (Mallat, 2008, section 4.3). In the case of a sine wave, we have $\boldsymbol{x}_\mathbf{a}(t) = \cos(2\pi t) + \mathrm{i} \times \sin(2\pi t) = \exp(2\pi \mathrm{i} t)$, of which we derive $|\boldsymbol{x}_\mathbf{a}|(t) = 1$. Provided that $\ddot{\boldsymbol{\tau}}(t) \ll 1$, the nonstationary wave $\mathcal{W}_\tau \boldsymbol{x}(t)$ is said to be asymptotic, and its analytic part is approximately equal to $\dot{\boldsymbol{\tau}}(t) \times \exp(2\pi \mathrm{i} \boldsymbol{\tau}(t))$.

Analytic wavelets

We denote by H^2 the space of analytic signals, that is, finite-energy functions whose Fourier transform vanishes over the half line of negative frequencies:

$$H^2 = \left\{ \boldsymbol{h}(t) \in L^2(\mathbb{R}, \mathbb{C}) \;\middle|\; \forall \omega < 0, \hat{\boldsymbol{h}}(\omega) = 0 \right\}.$$

The space H^2 is a subspace of $L^2(\mathbb{R}, \mathbb{C})$, closed under the action of shifts and dilations. Consequently, all dilations of an analytic wavelet $\boldsymbol{\psi}(t) \in H^2$ are themselves analytic. Moreover, since a convolution in the time domain is equivalent to a product in the Fourier domain, the result of the convolution $(\boldsymbol{x} * \boldsymbol{\psi})(t)$ between a real signal $\boldsymbol{x}(t) \in L^2(\mathbb{R}, \mathbb{R})$ and an analytic wavelet $\boldsymbol{\psi}(t) \in H^2$ is analytic. Therefore, it is judicious to analyze real signals with analytic wavelets,

hence mapping them to the space H^2 and subsequently demodulating oscillations with complex modulus.

Wavelets are able to track nonlinear time warps only if their time-frequency localization is sufficient to approximate the instantaneous frequency $\dot{\tau}(t)$ by a constant over the temporal support of every wavelet.

Scalogram ridges

Delprat et al. (1992) have proven that the Gabor scalogram extracts the instantaneous frequency $\dot{\tau}(t)$ of a chirp $\mathcal{W}_\tau x(t) = \dot{\tau}(t)\cos(2\pi\tau(t))$, even in the presence of small additive noise or other nonstationary components at other locations in the time-frequency domain:

$$\mathbf{U}_1(\mathcal{W}_\tau x)(t,\gamma) \approx \frac{1}{2}\dot{\tau}(t) \times \widehat{g_T}\left(2^\gamma - \dot{\tau}(t)\right),$$

where the approximation is valid for a chirp $\mathcal{W}_\tau x(t)$ whose instantaneous frequency has slow relative variations, i.e. satisfying the asymptoticity condition $\ddot{\tau}(t)/\dot{\tau}(t) \ll 1$. A more general result holds for a sine wave that is modulated both in amplitude and in frequency, i.e. of the form $x(t) = \alpha(t)\cos(2\pi\theta(t))$ where $\alpha(t)$ (resp. $\theta(t)$) is the instantaneous amplitude (resp. phase) of $x(t)$. The asymptotic regime assumes that $\alpha(t)$ and $\dot{\theta}(t)$ evolve slowly, as expressed by the conditions

$$\frac{\dot{\alpha}(t)}{\alpha(t)} \ll \frac{1}{\dot{\theta}(t)} \ll \frac{\dot{\theta}(t)}{\ddot{\theta}(t)}.$$

The inequality on the left implies that the instantaneous amplitude $\alpha(t)$ has slow relative variations over the duration of a pseudo-period $1/\dot{\theta}(t)$, while the inequality on the right guarantees that the pseudo-period itself is slowly varying (Flandrin, 2001). With both these conditions satisfied, Delprat et al. (1992) perform Taylor expansions of $\alpha(t)$ and $\theta(t)$, yielding approximately

$$\mathbf{U}_1 x(t,\gamma)(t) \approx \frac{1}{2}\alpha(t) \times \widehat{g_T}\left(2^\gamma - \dot{\theta}(t)\right) \tag{2.15}$$

as the dominant term. Their proof relies on the stationary phase principle, which states that most of the contribution to the scalogram is located in the vicinity of the time-frequency point(s) where the instantaneous frequencies of the signal and the analyzing wavelet cancel each other.

Like in the Hilbert-based definition of instantaneous amplitude (see Equation 2.14), the purpose of complex modulus is to demodulate oscillations brought by the convolution with the analytic wavelets.

It stems from the above that tracking the local maxima of the scalogram over the curvilinear ridge parameterized by $t \mapsto (t, \log_2 \dot{\theta}(t))$ enables to read the instantaneous amplitude $\alpha(t)$.

As a matter of fact, any real signal $\boldsymbol{x}(t)$ can be written as in an unique way as $\boldsymbol{\alpha}(t)\cos(2\pi\boldsymbol{\theta}(t))$, by defining its analytic amplitude as $\boldsymbol{\alpha}(t) = |\boldsymbol{x}_{\mathrm{a}}|(t)$ and its analytic phase as $\boldsymbol{\theta}(t) = \frac{1}{2\pi}\arg \boldsymbol{x}_{\mathrm{a}}(t)$, but this decomposition is rarely useful in practice because natural sounds are made of multiple sinusoidal components as well as unpitched impulses, thus failing to comply with the assumption of asymptoticity. Nevertheless, restricting a natural sound to a well-localized region of the time-frequency plane yields approximate asymptoticity. Again, this remark highlights the importance of time-frequency localization in the construction of analytic wavelets.

Shape from texture

Besides chirps, time warps can approximate a continuous space of sustained sounds caused by many gestural interactions, including rubbing, scratching, and rolling (Conan et al., 2013). Within a probabilistic framework, these sounds are modeled as

$$\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{X}(t) = \dot{\boldsymbol{\tau}}(t)(\boldsymbol{X}\circ\boldsymbol{\tau})(t),$$

where the stationary process $\boldsymbol{X}(t)$ is the response of a resonating material and $\dot{\boldsymbol{\tau}}(t)$ accounts for gestural dynamics, e.g. speed of friction. Recovering $\boldsymbol{\tau}(t)$ from a single realization of the locally dilated process $\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{X}(t)$ is an inverse problem named "shape from texture" (Malik and Rosenholtz, 1997), with applications in human-computer interaction and sound design (Thoret et al., 2014).

2.2.4 *Time reversal*

The time reversal operator, that is, $\boldsymbol{\tau}(t) = -t$, is simple to define and interpret. In the particular case of a sine wave $\boldsymbol{x}(t) = \cos(2\pi t + \varphi)$, time reversal merely entails a phase shift of 2φ, because $\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{x}(t) = \cos(2\pi\times(-t)+\varphi) = \cos(2\pi t - \varphi)$. Time reversal is thus not perceived for sine waves or superpositions thereof. However, due to the causality of damping in vibrating bodies, natural sounds typically follow an asymmetric envelope in the time domain, starting with a sharp onset and ending with a slower decay. Therefore, the reversal of an audio recording is noticeable at the time scale of a full acoustic event, such as a musical note or a spoken word, especially in the vicinity of onsets. Incidentally, short-term time reversal was recently applied to produce digital audio effects (Kim and Smith, 2014, 2015).

Time reversal of a real signal $\boldsymbol{x}(t)$ is equivalent to the complex conjugation of its Fourier transform $\widehat{\boldsymbol{x}}(\omega)$. As a consequence, the Fourier transform modulus $|\widehat{\boldsymbol{x}}(\omega)|$ is not only invariant to translation, but also invariant to time reversal. Yet, although invariance to translation is needed for classification, invariance to time reversal is an undesirable property.

The situation is different in the time-frequency domain. If the wavelet $\boldsymbol{\psi}(t)$ has a symmetric envelope, that is, if there exists a phase shift $\varphi \in [0; 2\pi[$ such that $\boldsymbol{\psi}(-t) = \mathrm{e}^{\mathrm{i}\varphi}\boldsymbol{\psi}(t)$, the wavelet transform is covariant with respect to time reversal:

$$\begin{aligned}
\mathbf{U}_1(\mathcal{W}_\tau \boldsymbol{x})(t,\gamma) &= \left|\mathrm{e}^{\mathrm{i}\varphi}\int_{-\infty}^{+\infty} \boldsymbol{x}(-t')\boldsymbol{\psi}_{2^\gamma,T(2^\gamma)}(t'-t)\,\mathrm{d}t'\right| \\
&= \left|\mathrm{e}^{\mathrm{i}\varphi}\right| \times \left|\int_{-\infty}^{+\infty} \boldsymbol{x}(t')\boldsymbol{\psi}_{2^\gamma,T(2^\gamma)}(-t-t')\,\mathrm{d}t'\right| \\
&= \mathbf{U}_1\boldsymbol{x}(t,\gamma).
\end{aligned}$$

Gabor wavelets have a symmetric envelope, as they satisfy $\boldsymbol{\psi}(-t) = -\boldsymbol{\psi}(t)$, i.e. $\varphi = \pi$. This is due to the fact that Gaussians have bell-shaped symmetric envelopes. Therefore, the Gabor wavelet transform commutes with time reversal:

$$\mathbf{U}_1(\mathcal{W}_\tau \boldsymbol{x})(t,\gamma) = \mathbf{U}_1\boldsymbol{x}(-t,\gamma).$$

A simple way to break this undesirable invariance is to choose $\boldsymbol{\psi}(t)$ as an asymmetric wavelet instead of a Gabor symmetric wavelet. In this regard, the Gammatone wavelet, shown in Figure 2.3, is a natural candidate. Section 4.5 will show that Gammatone wavelets outperform Gabor wavelets in a task of musical instrument classification in solo phrases.

Gammatone auditory filter

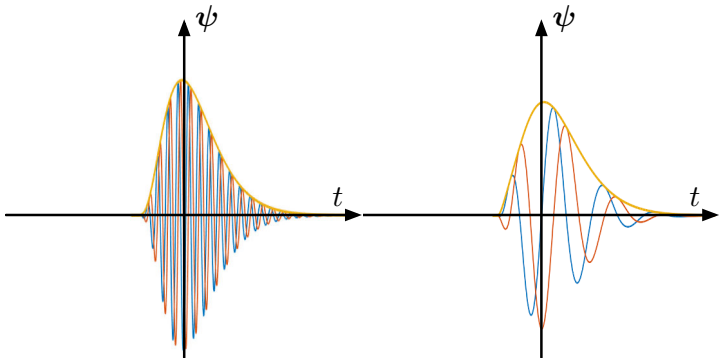
The complex-valued Gammatone wavelet is a modification of the real-valued Gammatone auditory filter, originated in auditory physiology. The Gammatone auditory filter of dimensionless frequency 1 is defined as a gamma distribution of order $N \in \mathbb{N}^*$ and bandwidth σ modulated by a sine wave, that is,

$$t^{N-1}\exp(-2\pi\sigma t)\cos(2\pi t).$$

For a fixed σ, the integer N controls the relative shape of the envelope, becoming less skewed as N increases. Psychoacoustical experiments have shown that, for $N = 4$, the Gammatone function provides a valid approximation of the basilar membrane response in the mammalian cochlea (Flanagan, 1960; Lyon, Katsiamis, and Drakakis, 2010; Patterson, 1976). In particular, it is asymmetric both in the time domain and in the Fourier domain, which allows to reproduce the asymmetry of temporal masking as well as the asymmetry of spectral masking (Fastl and Zwicker, 2007). It is thus used in computational models for auditory physiology (Pressnitzer and Gnansia, 2005). However, it does not comply with the Grossman-Morlet admissibility condition, because it has a non-negligible average. In addition, because the Gammatone auditory filter takes real values in the time domain, its Fourier transform satisfies Hermitian symmetry, which implies that

it does not belong to the space H^2 of analytic functions. More generally, there are no real-valued functions in H^2 (Grossmann and Morlet, 1984).

Pseudo-analytic Gammatone wavelet



(a) Gammatone wavelet, $Q = 8$, time domain

(b) Gammatone wavelet, $Q = 1$, time domain.

Figure 2.3: Pseudo-analytic Gammatone wavelets with different quality factors: (a) $Q = 8$, time domain; (b) $Q = 1$, time domain. Blue and red oscillations respectively denote real and imaginary parts, while the orange envelope denotes complex modulus.

With the aim of building a pseudo-analytic admissible Gammatone wavelet, Venkitaraman, Adiga, and Seelamantula (2014) have modified the definition of the Gammatone auditory filter, by replacing the real-valued sine wave $\cos(2\pi t)$ by its analytic part $\exp(2\pi i t)$ and by taking the first derivative of the gamma distribution, thus ensuring null mean. The definition of the Gammatone wavelet becomes

$$\boldsymbol{\psi}(t) = \left(2\pi(\mathrm{i}-\sigma)t^{N-1} + (N-1)t^{N-2}\right)\exp(-2\pi\sigma t)\exp(2\pi i t)$$

in the time domain, and

$$\widehat{\boldsymbol{\psi}}(\omega) = \frac{\mathrm{i}\omega \times (N-1)!}{(\sigma + \mathrm{i}(\omega-\sigma))^N}$$

in the Fourier domain. Besides its biological plausibility, the Gammatone wavelet enjoys a near-optimal time-frequency localization with respect to the Heisenberg uncertainty principle. Furthermore, this time-frequency localization tends to optimality as N approaches infinity, because the limit $N \to +\infty$ yields a Gabor wavelet (Cohen, 1995). Last but not least, the Gammatone wavelet transform of finite order N is causal, as opposed to the Morlet wavelet transform, which makes it better suited to real-time applications. From an evolutionary point of view, it has been argued that the Gammatone reaches a practical compromise between time-frequency localization and causality constraints. Venkitaraman, Adiga, and Seelamantula (2014) did not

provide a formula for deducing σ from the specification of a quality factor Q. We fill this gap in Appendix A.

The next section leaves the algebraic group of temporal diffeomorphisms $\mathcal{W}_\tau$ to consider a broader set of sound transformations that are expressed in the time-frequency domain and have no direct equivalent in the time domain.

2.3 ACOUSTICAL TRANSFORMATIONS

Many factors of variability between natural sounds are not amenable to mere diffeomorphisms of the time variable. Indeed, they may convey the sensation to transpose frequencies without affecting the time dimension; conversely, to stretch the time axis while leaving instantaneous frequencies unchanged; or, in full generality, to deform multiple components in different, albeit coherent, ways.

In this section, we define three kinds of transformations in the time-frequency domain that are ill-defined in the time domain: frequency transposition $\mathcal{F}$, time stretching $\mathcal{S}$, and deformation of the spectral envelope $\mathcal{E}$. Unlike time warps $\mathcal{W}$, these acoustical transformations affect either, but not both, the time or the frequency dimension. We review the state of the art on the topic, notably the phase vocoder, partial tracking, dynamic time warping (DTW), spectral shape descriptors, and mel-frequency cepstral coefficients (MFCC).

2.3.1 *Frequency transposition*

In this subsection, we define the notion of frequency transposition $\mathcal{F}_b$ as a translation by $b \in \mathbb{R}$ over the vertical axis γ of the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$. Although frequency transposition is ill-defined in the time domain, we give a closed-form expression of its effect on an amplitude-modulated additive sinusoidal model. We review the phase vocoder, a tool to re-synthesize a sound after frequency transposition in the time-frequency domain. We review partial tracking algorithms, an estimation procedure which aims at retrieving jointly the instantaneous frequencies $\dot{\boldsymbol{\theta}}_p(t)$ as well as their respective amplitudes $\boldsymbol{\alpha}_p(t)$ in an additive sinusoidal model. We formulate the "shape from texture" problem associated to frequency transposition.

Definition

The sensation of motion along the frequency axis is found in a wide variety of audio streams, including broadband environmental sounds or inharmonic tones. It should be noted that there is no need to postulate the existence of a fundamental frequency in the signal to define frequency transposition. Rather, as soon as the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ contains spectral cues as well as temporal cues, the notion of fre-

quency transposition arises and is distinguished from a mere time warp.

Frequency transposition is ill-defined in the signal domain. It is intuitively understood as a translation over the log-frequency axis in the scalogram, that is

$$\mathcal{F}_b(\mathbf{U}_1\boldsymbol{x})(t,\gamma) = \mathbf{U}_1\boldsymbol{x}(t,\gamma+b), \tag{2.16}$$

where the pitch interval b between $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ and it transposed version is measured in octaves. Observe that the time dimension in $\mathbf{U}_1\boldsymbol{x}$ is not affected by the acoustical transformation $\mathcal{F}_b$, whereas a geometrical time warp $\mathcal{W}_\tau$ would map (t,γ) to $(\boldsymbol{\tau}(t),\gamma+\log_2\dot{\boldsymbol{\tau}}(t))$.

However, the two-dimensional function $\mathcal{F}_b(\mathbf{U}_1\boldsymbol{x})(t,\gamma)$ is, generally speaking, not the wavelet scalogram of any signal. This is due to the strong redundancy of the continuous wavelet transform, whose time-frequency atoms overlap in the Fourier domain for neighboring values of the log-scale parameter γ. A necessary and sufficient condition for the transposed signal to exist is the reproducing kernel equation (Mallat, 2008, section 4.3), which is highly constraining.

Additive sinusoidal model

Despite the fact that there is no general formula that could yield $(\mathcal{F}_b\boldsymbol{x})(t)$ from any $\boldsymbol{x}(t)$, frequency transposition can be defined in closed form if we model $\boldsymbol{x}(t)$ as a sum of quasi-asymptotic components. The synthesis of sounds by stacking partial waves carrying independent amplitude modulations goes back to the early years of computer music (Risset, 1965).

Given an additive sinusoidal model of P partial waves

$$\boldsymbol{x}(t) = \sum_{p=1}^{P} \boldsymbol{\alpha}_p(t)\cos(2\pi\xi_p t), \tag{2.17}$$

and a pitch interval b in octaves, the signal

$$(\mathcal{F}_b\boldsymbol{x})(t) = \sum_{p=1}^{P} \boldsymbol{\alpha}_p(t)\cos(2\pi 2^b\xi_p t)$$

corresponds to a frequency transposition of $\boldsymbol{x}(t)$. Indeed, assuming that the partials ξ_p are segregated by the auditory filterbank, the scalogram of $(\mathcal{F}_b\boldsymbol{x})$ satisfies

$$\mathbf{U}_1(\mathcal{F}_b\boldsymbol{x})(t,\gamma) = \mathbf{U}_1\boldsymbol{x}(t,\gamma+b).$$

However, dilating $\boldsymbol{x}(t)$ with $\boldsymbol{\tau}(t) = 2^b t$ would yield

$$\mathcal{W}_\tau\boldsymbol{x}(t) = 2^b \times \sum_{p=1}^{P} \boldsymbol{\alpha}_p(2^b t)\cos(2\pi 2^b\xi_p t)$$

instead. Observe that, the rates of amplitude modulations $\boldsymbol{\alpha}_p(t)$ are dilated under the effect of the time warp $\mathcal{W}_\tau$, whereas they remain unchanged under the effect of frequency transposition $\mathcal{F}_b$. The difference is particularly noticeable in transient regions of the signal, wherein the relative variations of $\boldsymbol{\alpha}_p(t)$ are not negligible with respect to the time scale $T(2^\gamma)$.

Phase vocoder

Manipulating the time and frequency dimensions independently is the central purpose of the phase vocoder (Flanagan and Golden, 1966), a analysis-synthesis method based on the short-term Fourier transform (STFT) with widespread applications in computer music (Wishart, 1988) and speech processing. The realism of the phase vocoder in transient regions has been progressively improved by *ad hoc* alignment methods (Röbel, 2003). We refer to Liuni and Röbel (2013) for a recent survey.

In order to produce fine-grain sound manipulations with few artifacts without having to resort to such *ad hoc* alignment, a wavelet-based phase vocoder was developed by Kronland-Martinet (1988). However, when it comes to computational efficiency and numerical precision, implementing a multiresolution scheme for the auditory wavelet transform is a difficult problem in software engineering (Schörkhuber and Klapuri, 2010; Schörkhuber, Klapuri, and Sontacchi, 2013). This is why phase vocoders were originally designed with a short-term Fourier transform, which benefits from the simplicity of a single-resolution discrete scheme.

Partial tracking

A different problem in which frequency transposition plays a crucial role is partial tracking McAulay and Quatieri (1986), an estimation procedure which aims at retrieving jointly the instantaneous frequencies $\dot{\boldsymbol{\theta}}_p(t)$ as well as their respective amplitudes $\boldsymbol{\alpha}_p(t)$ in an additive sinusoidal model of the form

$$\boldsymbol{x}(t) = \sum_{p=1}^{P} \boldsymbol{\alpha}_{\mathrm{p}}(t) \cos(2\pi \boldsymbol{\theta}_p(t)).$$

Partial tracking algorithms extract local magnitude peaks of the short-term Fourier transform, which are subsequently linked through time according to continuity priors in $\dot{\boldsymbol{\theta}}_p(t)$ and $\boldsymbol{\alpha}_{\mathrm{p}}(t)$.

Because it is a bottom-up approach with a relatively short temporal context, partial tracking is prone to false detection in the presence of noise, despite recent enhancements in the prediction stage (Kereliuk and Depalle, 2008; Lagrange, Marchand, and Rault, 2007). Therefore, although the notion of frequency transposition remains an important

form of variability among natural sounds, modern audio classification systems do not strive to track partials as a pre-processing step. Instead, they adopt a holistic approach to pattern matching, with little or no prior detection, segmentation, or denoising.

Probabilistic outlook

Besides additive models, frequency transposition has recently been studied within a probabilistic framework, by means of the continuous wavelet transform (Omer and Torrésani, 2016). For $\boldsymbol{X}(t)$ a stationary process and $\boldsymbol{b}(t)$ a slowly varying signal, the "shape from texture" problem associated to frequency transposition consists in estimating both from a single realization $\boldsymbol{x}(t)$ of $\mathcal{F}_{\boldsymbol{b}}\boldsymbol{X}(t)$, whose scalogram $\mathbf{U}_1\boldsymbol{x}(t)$ is assumed to satisfy the identity

$$\mathcal{F}_{\boldsymbol{b}}(\mathbf{U}_1\boldsymbol{x})(t,\gamma) = \mathbf{U}_1\boldsymbol{x}(t,\gamma+\boldsymbol{b}(t)).$$

Mapping sounds to the time-frequency domain allows to study a non-stationary signal $\mathcal{F}_{\boldsymbol{b}}\boldsymbol{x}$ as the composition of two simpler functions, a Gaussian noise $\boldsymbol{x}(t)$ and a slow deterministic deformation $\boldsymbol{b}(t)$. Once again, the continuous wavelet transform reveals a form of regularity that was not readily available in the time domain.

2.3.2 *Time stretching*

Within a piece of classical music, even if the melodic contour is determined by the composer, the choice of rhythmic interpretation is left, up to some extent, to the performers. Likewise, the flow rate of phonetic units in continuous speech may vary across utterances of the same text, due to speaker or mood variability. Both of these effects convey the impression to stretch the time axis. Therefore, audio features for classification should remain robust to time stretches while providing an insight on fine-scale temporal structure.

In this subsection, we address the problem of modeling the effects of time stretching in the time-frequency domain. We show how the phase vocoder, introduced in Subsection 2.3.1, is suitable for analysis-synthesis of stretched sounds. We discuss the limitations of dynamic time warping (DTW) for time series alignment of short-term audio descriptors. We round off the subsection with a forward reference to the temporal scattering transform, found in Section 3.1.

Definition

Time stretching can be formulated in the time-frequency domain as a transformation $\mathcal{S}_\tau$ on the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$, where the diffeomorphism $\boldsymbol{\tau}(t)$ warps the time variable t without affecting the log-frequency variable γ, thus following the equation

$$\mathcal{S}_\tau(\mathbf{U}_1\boldsymbol{x})(t,\gamma) = \mathbf{U}_1\boldsymbol{x}(\boldsymbol{\tau}(t),\gamma).$$

Again, the existence of a signal $(\mathcal{S}_\tau \boldsymbol{x})(t)$ whose continuous wavelet transform would be $\mathcal{S}_\tau(\mathbf{U}_1\boldsymbol{x})(t,\gamma)$ is conditional upon the reproducing kernel equation, which is highly constraining — see Mallat (2008, section 4.3). Therefore, in spite of its simple definition in the time-frequency domain, time stretching is ill-defined in the time domain.

Phase vocoder for time stretching

Observe that, in the asymptotic regime, frequency transposition and time stretching are dual to each other. Indeed, the composition of these two operators boils down to a time warp $\mathcal{W}_\tau$ in the time domain:

$$\begin{aligned}(\mathcal{F}_{\log_2 \dot{\tau}} \circ \mathcal{S}_\tau)(\mathbf{U}_1\boldsymbol{x})(t,\gamma) &= \mathbf{U}_1\boldsymbol{x}(\tau(t), \gamma + \log_2 \dot{\tau}(t)) \\ &= \mathbf{U}_1(\mathcal{W}_\tau \boldsymbol{x})(t,\gamma).\end{aligned}$$

Consequently, composing a frequency transposition $\mathcal{F}_{-\log_2 \dot{\tau}}$ with a time warp $\mathcal{W}_\tau$ yields a time stretching $\mathcal{S}_\tau$:

$$\begin{aligned}\mathcal{F}_{-\log_2 \dot{\tau}}(\mathbf{U}_1\mathcal{W}_\tau\boldsymbol{x})(t,\gamma) &= \mathbf{U}_1\boldsymbol{x}(\tau(t),\gamma) \\ &= \mathcal{S}_\tau(\mathbf{U}_1\boldsymbol{x})(t,\gamma)\end{aligned}$$

Drawing on the above, the phase vocoder is suited to both frequency transposition and time stretching.

Dynamic time warping

Building stretch-invariant similarity metrics for pattern matching is at the heart of dynamic time warping (DTW) algorithms (Sakoe and Chiba, 1978), which rely on dynamic programming heuristics to find a stretch $\tau(t)$ which approximately aligns a query $\mathbf{U}\boldsymbol{x}(t,\gamma)$ to a reference template $\mathbf{U}\boldsymbol{y}(t,\gamma)$. In discrete time, so as to mitigate the combinatorial explosion of possible alignments, the amount of deviation is constrained to be lower than some constant T, i.e.

$$|\tau(t) - t| < T, \tag{2.18}$$

where T is typically set to about 10% of the duration of the query $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$. For example, dynamic time warping has been used successfully to classify vocalizations of marine mammals (Brown and Miller, 2007). We refer to the textbook of Müller (2007, chapter 4) for an overview of dynamic time warping in music signal analysis.

Like partial tracking, dynamic time warping has the downside of being prone to misalignment, because it relies on strong assumptions on the observed data. If the query $\mathbf{U}\boldsymbol{x}(t,\gamma)$ no longer satisfies the assumption of being amenable to a reference template $\mathbf{U}\boldsymbol{y}(t,\gamma)$ by some hypothetical time stretch $\tau(t)$, the behavior of dynamic time warping is ill-defined. Therefore, dynamic time warping is not appropriate for

long audio streams with overlapping sources, such as acoustic scenes or polyphonic music.

Drawing a parallel with Subsection 2.2.1, the constant T in Equation 2.18 can be interpreted as an amount of invariance with respect to time shifts and local time warps. In audio classification, even in cases when explicit dynamic time warping is not conceivable, accounting for time shifts and time warps in the representation is of utmost importance. Again, analytic wavelets play a central role in this matter — see Subsection 2.2.3. The temporal scattering transform, which will be formally introduced in Section 3.1, is naturally derived from this observation, as it consists of a two-layer cascade of wavelet transforms interspersed with modulus nonlinearities, thus capturing time warps in the time domain as well as time stretches in the time-frequency domain.

2.3.3 *Variations in spectral envelope*

In this subsection, we model variations of the spectral envelope by means of a diffeomorphism $\mathcal{E}_\tau$ of the frequency variable operating over the scalogram. We discuss the limitations of spectral shape descriptors based on high-order statistical moments, such as spectral skewness and kurtosis. We review the construction of the power cepstrum and the mel-frequency cepstrum.

Definition

Any voiced sound in its steady state, such as a musical note or a spoken phoneme, can locally be approximated by an additive model, in which each sinusoid appears as a peak in the short-term Fourier spectrum. The spectral envelope is intuitively understood as a smooth curve interpolating the locations and magnitudes of these peaks. From an acoustical perspective, the spectral envelope accounts for the modes of vibration of the resonator, that is, the vocal tract in speech or the body of the instrument in music. The purpose of short-term audio descriptors for classification is to characterize the overall shape of the spectral envelope while remaining stable to small changes in the locations and magnitudes of spectral peaks.

In analogy with the previous definitions of time warps $\mathcal{W}_\tau$, frequency transpositions $\mathcal{F}_b$, and time stretches $\mathcal{S}_\tau$, deformations of the spectral envelope $\mathcal{E}_\tau$ can be formulated by applying a warp $\boldsymbol{\tau}(\gamma)$ to the log-frequency dimension in a scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ while leaving the time dimension unchanged:

$$\mathcal{E}_\tau \mathbf{U}_1 \boldsymbol{x}(t,\gamma) = \dot{\boldsymbol{\tau}}(\gamma) \times \mathbf{U}_1 \boldsymbol{x}(t, \boldsymbol{\tau}(\gamma)).$$

Again, it should be kept in mind that the deformed wavelet transform modulus $\mathcal{E}_\tau \mathbf{U}_1 \boldsymbol{x}(t,\gamma)$ in the time-frequency domain may not have an equivalent in the time domain.

Instability of spectral shape descriptors

A substantial part of the general-purpose audio descriptors for classification aim at characterizing the spectral envelope by regarding it as a statistical distribution which takes values in the Fourier domain. This is motivated by the fact that the Fourier transform magnitude $|\hat{\boldsymbol{x}}|^2(\omega)$ is nonnegative and that its sum is equal to the squared L^2 norm of $\boldsymbol{x}(t)$, in virtue of Plancherel's identity. Besides spectral centroid and spectral flatness, which respectively correspond to the mean and the variance of $|\hat{\boldsymbol{x}}|(\omega)$, higher-order moments, such as skewness and kurtosis, are often included to the collection of features for audio classification (Peeters, 2004).

An alternative point of view is that these statistical moments decompose the magnitude of the Fourier transform over an orthogonal basis of monomials of growing degree $k > 0$, i.e. $\int \omega^k |\hat{x}|(\omega)\ \mathrm{d}\omega$. Lastly, because the Fourier transform of $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t)$ is equal to $\mathrm{i}\omega\hat{\boldsymbol{x}}(\omega)$, they also correspond to the Fourier transform moduli of the k^{th} derivatives of $\boldsymbol{x}(t)$ in the time domain.

However, for high values of the k, the k^{th} statistical moment of $|\hat{\boldsymbol{x}}|(\omega)$ is excessively sensitive to small variations of the spectral envelope, and in particular to dilations by some constant $a \neq 1$, that is, to frequency transpositions — see Subsection 2.3.1. Indeed, a homothetic change of variable in the k^{th} statistical moment of $\frac{1}{a} \times |\widehat{\boldsymbol{x}}|(\frac{\omega}{a})$ yields the identity

$$\frac{1}{a} \times \int \omega^k |\widehat{\boldsymbol{x}}| \left(\frac{\omega}{a}\right) \mathrm{d}\omega \quad = \quad a^k \times \int \omega^k |\widehat{\boldsymbol{x}}|(\omega)\ \mathrm{d}\omega.$$

Consequently, after dilating the spectral envelope of $|\widehat{\boldsymbol{x}}|$ by a, the spectral centroid ($k = 1$) gets multiplied by a ; the spectral flatness ($k = 2$), by a^2 ; the spectral skewness ($k = 3$), by a^3, and so forth. Yet, robust features for audio classification should be stable to the dilation operator $|\widehat{\boldsymbol{x}}|(\omega) \mapsto a|\widehat{\boldsymbol{x}}|(a\omega)$, i.e. they should be affected linearly, instead of polynomially, by the amount a of frequency transposition (Andén and Mallat, 2014). The same instability property applies to a broader class of transformations of the spectral envelope, but its proof is more difficult to carry out in full generality.

Power cepstrum

In order to build a stabler description of the spectral envelope, the basis of monomials ω^k can be replaced by a Fourier basis $\exp(2\pi \mathrm{i} k\omega)$ with $k \in \mathbb{N}$, hence leading to the notion of *cepstrum*, which is ubiquitous in signal processing and in particular in speech recognition. The term "cepstrum" was coined by Tukey (Bogert, Healy, and Tukey, 1963) as an anagram of the word "spectrum". Indeed, it operates on the frequency variable ω in ways customary of the time variable t, and vice versa (Oppenheim and Schafer, 2004).

There are multiple mathematical definitions associated to the notion of cepstrum, including complex cepstrum, real cepstrum, and phase cepstrum. The definition used in audio classification is named power cepstrum, as it consists in the Fourier transform modulus of the logarithm of the spectral envelope, i.e.

$$\text{cepstrum}(\boldsymbol{x})(k) = \left| \int_{-\infty}^{+\infty} \log |\widehat{\boldsymbol{x}}(\omega)| \exp(-2\pi \mathrm{i} k\omega) \, \mathrm{d}\omega \right|.$$

Let $\boldsymbol{x}(t) = \sum_p \alpha_p \cos(2\pi p\xi t)$ be a harmonic additive signal of fundamental frequency ξ. The envelope of the Fourier transform of $\boldsymbol{x}(t)$ is $|\widehat{\boldsymbol{x}}|(\omega) = \sum_p \alpha_p \times \delta(\omega - p\xi)$, i.e. a sequence of bumps of period ξ. By the Poisson summation formula, the power cepstrum of $\boldsymbol{x}(t)$ is a sequence of bumps of period $1/\xi$. The quantity $1/\xi$ is homogeneous to the inverse of a frequency, that is, to a temporal duration; it is called a *quefrency*, by anagram with frequency. Because the sequence of log-amplitudes $(\log \alpha_p)_p$ has slow variations according to the partial wave index p, the cepstral bump of low quefrency $1/\xi$ dominates higher-quefrency bumps in the cepstrum of $\boldsymbol{x}(t)$. Therefore, the period $1/\xi$ of $\boldsymbol{x}(t)$ can be measured by finding the global maximum of its Fourier cepstrum (Noll and Schroeder, 1967). This measure is more robust to inharmonic deviations from exact multiples $p\xi$, and is robust to the removal of the fundamental partial wave, i.e. to setting $\alpha_1 = 0$.

The Fourier power cepstrum of a harmonic sound with a missing fundamental is shown in Figure 2.4. Interestingly, cepstral transformations remain at the core of state-of-the-art algorithms for fundamental frequency estimation to this date, although the Fourier spectrum $|\widehat{\boldsymbol{x}}|(\omega)$ is replaced by autocorrelation (De Cheveigné and Kawahara, 2002; Mauch and Dixon, 2014) in order to avoid octave errors.

Mel-frequency cepstral coefficients

The most widespread short-term descriptor for audio classification is the mel-frequency cepstrum (MFC), a variant of the power cepstrum with two major differences. First, it is based on the mel scale of frequencies γ, which is roughly linear below 1000 Hz and logarithmic above, instead of the linear scale ω. Secondly, is is obtained by a cosine transform instead of the complex modulus of a Fourier transform. This is in order to yield descriptors that are approximately of null average through time, thus complying better with the assumption of Gaussian normality of features, commonly made in machine learning.

The definition of the mel-frequency cepstrum is

$$\text{MFC}(\boldsymbol{x})(t,k) \quad = \quad \int_{-\infty}^{+\infty} \log \mathbf{U}_1 \boldsymbol{x}(t,\gamma') \times \cos(2\pi k\gamma') \, \mathrm{d}\gamma'.$$

The mel-frequency cepstrum is a two-dimensional array indexed by time t and quefrency k. High quefrencies k correspond to fine details

of the scalogram, such as the time-frequency edges produced by the presence of harmonic partials, whereas low quefrencies k capture the spectral envelope (Davis and Mermelstein, 1980).

Let $\mathcal{E}_\tau$ be a deformation of the spectral envelope. In the cepstral domain, the special case $\tau(\gamma) = a\gamma$ causes a contraction of the cepstrum by the inverse factor $\frac{1}{a}$ over the variable k, as proven by the following change of variable:

$$\begin{aligned}
\mathrm{MFC}(\mathcal{E}_\tau(\mathbf{U}_1 x))(t,k) &= a \times \int_{-\infty}^{+\infty} \log \mathbf{U}_1 x(t, a\gamma') \times \cos(2\pi k \gamma')\, \mathrm{d}\gamma' \\
&= \int_{-\infty}^{+\infty} \log \mathbf{U}_1 x(t, \gamma') \times \cos\left(2\pi \frac{k}{a}\gamma'\right)\, \mathrm{d}\gamma' \\
&= \mathrm{MFC}(\mathbf{U}_1 x)\left(t, \frac{k}{a}\right).
\end{aligned}$$

It appears from the above that the k^{th} coefficient in $\mathrm{MFC}(x)(t,k)$ is roughly invariant to the action of $\tau(\gamma) = a\gamma$ as long as $k \ll a$. Within a discrete setting, the mel-frequency cepstrum is computed with a type-II discrete cosine transform, and the index k takes integer values. A common practice in audio classification is to discretize γ with a bank of 40 auditory filters and then retain only the 12 lowest values of k (Eronen and Klapuri, 2000). First and second discrete-time derivatives of the mel-frequency cepstrum, known as "deltas" and "delta-deltas", are frequently appended to this 12-dimensional representation.

Despite their relative simplicity, MFCC features are surprisingly effective in audio-based similarity retrieval and classification. Since their earliest use in music instrument classification (Brown, 1999), many scientific publications have strived to refine incrementally the design of MFCC features, by tailoring them more specifically to musical audio. However, they appeared to yield little or no benefit with respect to a baseline of 12 MFCC features computed over short-term windows of size $T = 25\,\mathrm{ms}$, and aggregated into a Gaussian mixture model — see (Aucouturier and Pachet, 2004) for a critical review. As a result, this baseline is still in use to this date (Stowell, 2015).

DTW algorithm over the log-frequency axis

Whereas dynamic time warping of time-frequency representations is ubiquitous in audio signal processing — see e.g. Damoulas et al. (2010) in bioacoustics — few existing contributions explicitly account for the deformations of the log-frequency axis. We should, however, mention Hemery and Aucouturier (2015), who apply the DTW algorithm over the variable γ in a spectrotemporal receptive field (STRF) representation, in the context of unsupervised acoustic scene classification. The STRF representation will be discussed in detail in Section 3.2, devoted to time-frequency scattering.

2.4 TRANSFORMATIONS OF MUSICAL PARAMETERS

The two previous sections respectively addressed low-level transformations of audio waveforms and mid-level transformations in the time-frequency domain. At a higher level of abstraction, musicians refer to sound according to a variety of "parameters" (Meyer, 1989), such as pitch, timbre, and rhythm.

In this section, we strive to relate these high-level parameters to transformations in the time-frequency domain. To this aim, we define the stationary source-filter as a convolution between a harmonic excitation and a broadband filter. We define the pitch shift as an operator which transposes the source while leaving the filter unchanged. Then, we show that the discrete cosine transform involved in the computation of mel-frequency cepstral coefficients (MFCC) segregates frequency transposition from deformations of the spectral envelope. However, we provide empirical evidence demonstrating that MFCC are not as invariant to realistic pitch shifts as they are to changes in loudness, interpret, or instrument manufacturer. We conclude that the currently available timbral descriptors do not properly disentangle pitch from timbre, thus appealing for the use of machine learning techniques in musical instrument classification.

2.4.1 *Pitch*

In this subsection, we address the problem of building pitch-invariant representations for audio classification. We define pitch as the only attribute of music which is relative, intensive, and independent of instrumentation. We distinguish frequency transposition, a pure translation along the log-frequency, from a realistic pitch shift, which translates spectral peaks but not the spectral envelope. According to the stationary source-filter model, we show that mel-frequency cepstral coefficients (MFCC) disentangle the contributions of the source and the filter, and provide some invariance to frequency transposition. Yet, we provide empirical evidence to demonstrate that MFCC are not invariant to realistic pitch shifts.

Definition

Among the cognitive attributes of music, pitch is distinguished by a combination of three properties. First, it is relative: ordering pitches from low to high gives rise to intervals and melodic patterns. Secondly, it is intensive: multiple pitches heard simultaneously produce a chord, not a single unified tone — contrary to loudness, which adds up with the number of sources. Thirdly, it does not depend upon instrumentation, thus making possible the transcription of polyphonic music under a single symbolic system (Cheveigné, 2005).

Like the formal distinction between detection and classification, there is a formal distinction between pitch-covariant and pitch-invariant tasks in music information retrieval, respectively epitomized by score transcription and musical instrument classification.

It should be reminded that the notions of covariance and invariance to pitch shift are subordinated to the choice of a pitch range Γ over the log-frequency dimension γ, which plays the same role as the time scale T over the time dimension t. A desirable representation for musical instrument classification should be invariant to pitch shifts up to the tessitura Γ of instruments, that is, about two octaves. In contrast, classification of urban or environmental sounds requires a smaller amount of invariance.

Stationary source-filter model

The additive sinusoidal model defined in Equation 2.17 is made of a superposition of partial tones of frequencies $(\xi_p)_p$ and amplitudes $(\alpha_p)_p$. A pitch is perceived if all partial frequencies are multiples of a fundamental frequency ξ, that is, if $\xi_p = p\xi$. Once this constraint is added, the harmonic sinusoidal model is of the form

$$\boldsymbol{x}(t) = \sum_{p=1}^{P} \alpha_p \cos(2\pi p \xi t).$$

Before defining the notion of pitch shift for the additive model above, it is necessary to decompose $\boldsymbol{x}(t)$ into a convolution of two terms, respectively called source and filter. The source

$$\boldsymbol{e}(t) = \sum_{p=1}^{P} \cos(2\pi p \xi t)$$

is parameterized by the choice of fundamental frequency ξ and is not affected by the amplitudes $(\alpha_p)_p$. Conversely, the filter $\boldsymbol{h}(t)$ is defined in the Fourier domain as a smooth, spectral envelope which satisfies

$$|\widehat{\boldsymbol{h}}|(p\xi) = \alpha_p,$$

thus connecting the locations ξ_p and magnitudes α_p of spectral peaks. Once $\boldsymbol{e}(t)$ and $\boldsymbol{h}(t)$ have been defined, the stationary source-filter is the convolution

$$\begin{aligned} \boldsymbol{x}(t) &= (\boldsymbol{e} * \boldsymbol{h})(t) \\ &= \sum_{p=1}^{P} \hat{\boldsymbol{h}}(p\xi) \cos(2\pi p \xi t). \end{aligned}$$

In speech technology, the purpose of the source $\boldsymbol{e}(t)$ is to model the periodic, impulsive airflow emerging from the glottis, whereas the filter $\boldsymbol{h}(t)$ accounts for the resonance of the vocal tract. As long as

the rank of the partial is lower than the quality factor of the auditory filter bank, that is, typically $P < Q$, each wavelet in the auditory filter bank resonates to a most one partial. Consequently, the spectral envelope $|\widehat{\boldsymbol{h}}(2^{\gamma})|$, also known as formantic structure, may be factorized out from the finite sum of partials, yielding

$$\mathbf{U}_1\boldsymbol{x}(t,\gamma) = |\widehat{\boldsymbol{h}}|(2^{\gamma}) \times \mathbf{U}_1\boldsymbol{e}(t,\gamma).$$

The typical shapes of the source and the filter in the Fourier domain are shown in Figure 2.5.

The stationary source-filter model is historically linked with the development of the mel-frequency cepstrum for speech recognition — see Subsection 2.3.3. Indeed, taking the logarithm of both sides allows to convert this product into a sum:

$$\log \mathbf{U}\boldsymbol{x}(\gamma) = \log \mathbf{U}\boldsymbol{e}(\gamma) + \log \mathbf{U}\boldsymbol{h}(\gamma) \tag{2.19}$$

By linearity of the cosine transform, the additive property of the log-spectra is transferred to the cepstra

$$\mathrm{MFC}(\boldsymbol{x})(k) = \mathrm{MFC}(\boldsymbol{e})(k) + \mathrm{MFC}(\boldsymbol{h})(k).$$

Since the source $\boldsymbol{e}(t)$ is made of equidistant peaks in the Fourier domain, its scalogram $\mathbf{U}\boldsymbol{e}(\gamma)$ is an irregular function of the log-frequency variable γ. In turn, the spectral envelope of the filter $\boldsymbol{h}(t)$ is a regular function of γ. This notion of regularity can be linked to the rate of decay associated to the cepstral variable k (Mallat, 2008, Section 2.3). Henceforth, the cepstrum $\mathrm{MFC}(\boldsymbol{h})(k)$ of the filter $\boldsymbol{h}(t)$ has a faster decay than the cepstrum $\mathrm{MFC}(\boldsymbol{e})(k)$ of the source $\boldsymbol{e}(t)$. As a result, keeping only the lowest values of k in $\mathrm{MFC}(\boldsymbol{x})(k)$ yields a representation which is relatively invariant to deformations of the source $\boldsymbol{e}(t)$, such as speaker variability, while remaining relatively discriminative to deformations of the filter $\boldsymbol{h}(t)$, such as phonetic variability.

Preservation of formantic structure

The frequency transposition operator $\mathcal{F}_b$, introduced in Subsection 2.3.1, transposes spectral peaks $p\xi$ to new locations $2^b p\xi$ without affecting the amplitudes α_p of corresponding partials:

$$\mathcal{F}_b(\boldsymbol{x})(t) \quad = \quad \sum_{p=1}^{P} \alpha_p \cos(2\pi \times 2^b p\xi t).$$

In the time-frequency domain, applying $\mathcal{F}_b$ on the stationary source-filter model yields a translation of both the spectral peaks and the spectral envelope:

$$\mathcal{F}_b(\mathbf{U}_1\boldsymbol{x})(t,\gamma) = \widehat{\boldsymbol{h}}(2^{\gamma+b}) \times \mathbf{U}_1\boldsymbol{e}(t,\gamma+b).$$

In contrast, a realistic pitch shift $\mathcal{P}_b$ by a log-frequency interval b only translates the spectral peaks while leaving the spectral envelope unchanged:

$$\mathcal{P}_b(\mathbf{U}_1\boldsymbol{x})(t,\gamma) = \widehat{\boldsymbol{h}}(2^\gamma) \times \mathbf{U}_1\boldsymbol{e}(t,\gamma+b).$$

In the source-filter model, the exact value of the spectral envelope is only available at frequencies $p\xi$. Therefore, the definition above is not directly applicable in practice, because the amplitude $|\widehat{\boldsymbol{h}}|(2^b p\xi)$ of the spectral envelope at the frequency $2^b p\xi$ is generally not known. However, the overall shape of the spectral envelope can be obtained by interpolating the sequences of locations ξ_p and magnitudes α_p of spectral peaks. This interpolation is performed by low-pass *liftering* (anagram of filtering) over the log-frequency axis γ. It has led Röbel and Rodet (2005) to propose an improvement of the phase vocoder which preserves the formantic structure of frequency transposition.

Experimental validation

Two musical pitches played by the same instrument, other things being equal, are not exactly amenable to each other by a translation on the log-frequency axis γ. This is illustrated in Figure 2.6.

For small intervals, a frequency transposition of the additive sinusoidal model is sufficient to approximate a realistic pitch shift, and the residual terms remain small with respect to the translation term. However, large pitch intervals within the same instrument also affect the qualitative timbral properties of sound, which is not taken into account in the scalogram. The same property is verified for a wide class of musical instruments, especially brass and woodwinds.

How pitch-invariant is the mel-frequency cepstrum ?

From this observation, we argue that the construction of powerful invariants to musical pitch is insufficiently approximated by translation-invariant representations on the log-frequency axis, such as the discrete cosine transform (DCT) underlying the mel-frequency cepstrum — see Subsection 2.3.3. To validate this claim, we have extracted the mel-frequency cepstral coefficients (MFCC) of 1116 individual notes from the Real World Computing (RWC) dataset (Goto et al., 2003), as played by 6 instruments, with 32 pitches, 3 nuances, and 2 interprets and manufacturers. When more than 32 pitches were available (e.g. piano), we selected a contiguous subset of 32 pitches in the middle register. We then have compared the distribution of squared Euclidean distances between musical notes in the 12-dimensional space of MFCC features.

Figure 2.7 summarizes our results. We found that restricting the cluster to one nuance, one interpret, or one manufacturer hardly re-

duces intra-class distances. This suggests that MFCC are fairly successful in building invariant representations to such factors of variability. In contrast, the cluster corresponding to each instrument is shrunk if decomposed into a mixture of same-pitch clusters, sometimes by one order of magnitude. In other words, most of the variance in an instrument cluster of MFCC is due to pitch transposition. Keeping less than 12 coefficients certainly improves invariance, yet at the cost of inter-class variability between instruments, and vice versa.

2.4.2 *Timbre*

Building robust yet discriminative timbral descriptors is a crucial preliminary step in audio classification. In this subsection, we distinguish two definitions of timbre. The former refers to abstract properties of sound, whereas the latter is directly related with the identification of sources.

A negative definition of timbre

The notion of musical timbre is defined in a negative way, as the cognitive attribute which is left unchanged by variations in pitch, intensity, and duration. As such, it bears an equivocal meaning, depending on whether it refers to the properties of sound itself or to the source that emits it. Common parlance conflates the two definitions (Casati and Dokic, 1994), because identifying sources is arguably the foremost purpose of hearing from an evolutionary perspective. Nevertheless, distinguishing them carefully is crucial to the design of a machine listening system.

Timbre as a qualitative attribute

The first definition hypothesizes that our perception of timbre is based on a multidimensional, continuous representation (Siedenburg, Fujinaga, and McAdams, 2016). This hypothesis is put to trial by gathering a set of adjectives which are commonly employed to describe sound, e.g. dry, rough, warm, rich, soft, bright, round, and so forth. Observe that none of these adjectives belong exclusively to the vocabulary of audition ; instead, their intuitive signification arises by analogy with visual or tactile sensations (Faure, McAdams, and Nosulenko, 1996). Each of them can be readily employed as a rating criterion for a listening experiment over a controlled dataset of recorded sounds, be them instrumental, vocal, or environmental.

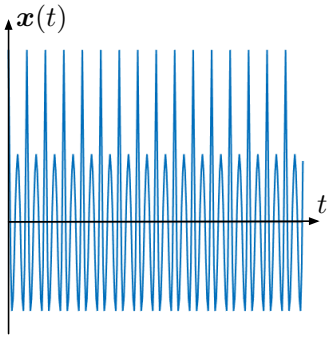
With this aim in mind, McAdams et al. (1995) have measured similarity ratings between pairs of isolated musical notes and applied multidimensional scaling (MDS), a linear dimensionality reduction technique, to recover an underlying mental representation of timbre. They found that, once obliterated inter-subject variability, timbral sim-

ilarity boils down to an Euclidean space of dimension 3, whose principal components roughly correlate with short-term audio descriptors, namely spectral centroid, spectral flux, and attack time — see Subsection 2.3.3. In other words, behind the wide variety of adjectives employed to rate sound qualities, many of them are near-synonyms (e.g. "round" and "warm"), and only a few actually relate to independent factors of variability. Yet, given a task of musical instrument recognition, human listeners significantly outperform automatic classifiers trained on the aforementioned three-dimensional feature space.

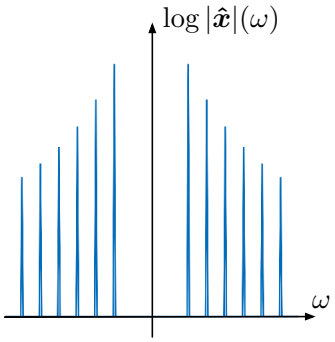
Timbre as a discriminative attribute

The lesson to be learned from this somewhat disappointing result is that the human ability of distinguishing sources is not merely a perceptual comparison in absolute terms, but rather an adaptive process which also takes into account other phenomena, such as pitch, intensity, phrasing dynamics, as well as priming effects. It follows from the above that, when speaking of a "bright timbre", the word "timbre" proceeds from a different definition than in the "timbre of the trumpet" (Castellengo and Dubois, 2007). The former, related to tone quality, expresses a property of sound, independently from the source that produced it. On the contrary, the latter, related to tone identity, is peculiar to the source and may encompass several timbral qualities in conjunction with other attributes.

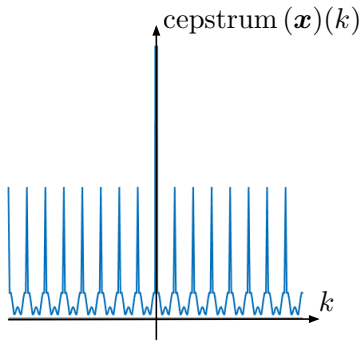
Owing to nonlinear wave propagation along the bore, trumpet sounds may convey various levels of brightness according to interpretation (Norman et al., 2010; Risset and Mathews, 1969). More generally, it seems that what music psychologists call the "bandwidth for timbre invariance" is of the order of one octave (Handel and Erickson, 2001), or even below (Steele and Williams, 2006). For an automatic classifier to overcome this limitation, the resort to machine learning, in addition to a powerful signal representation, appears as necessary.



(a) Time domain

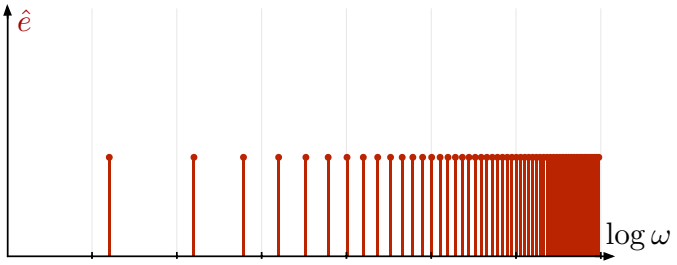


(b) Log-magnitude of the Fourier spectrum.

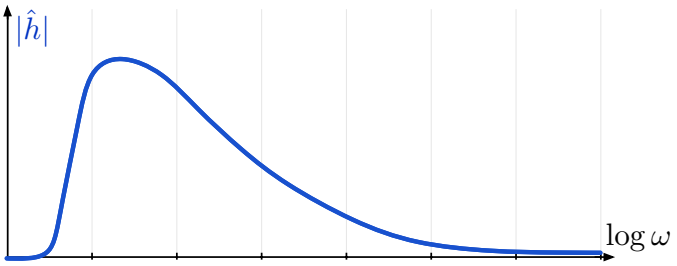


(c) Fourier power cepstrum.

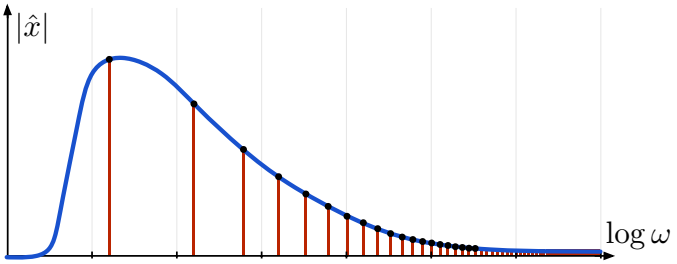
Figure 2.4: Retrieving the missing fundamental. Figure 2.4a shows a harmonic signal $\boldsymbol{x}(t)$ of fundamental frequency ξ in the time domain, built by stacking partial waves of frequencies 2ξ, 3ξ, 4ξ, etc. Figure 2.4b shows the log-magnitude $\log|\hat{\boldsymbol{x}}|(\omega)$ of the Fourier spectrum. Observe the missing fundamental ξ. Figure 2.4c shows the power cepstrum of $\boldsymbol{x}(t)$. Observe that the period ξ^{-1} is accurately retrieved.



(a) Spectrum of the source.

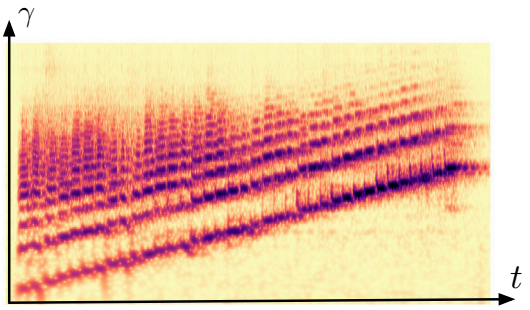


(b) Spectrum of the filter.

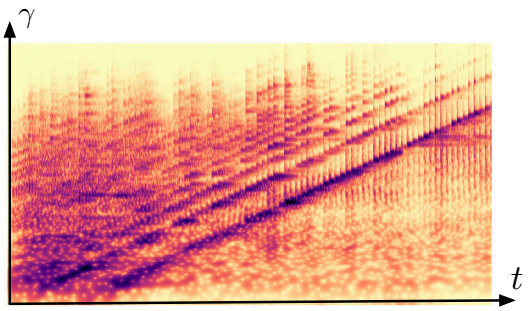


(c) Spectrum of the source-filter model.

Figure 2.5: The stationary source-filter model. The source is a superposition of partials of respective frequencies $p\xi$ for integer p. The filter is a broadband spectral envelope in the Fourier domain.



(a) Tuba chromatic scale.



(b) Piano chromatic scale.

Figure 2.6: Pitch shifts are not frequency transpositions.

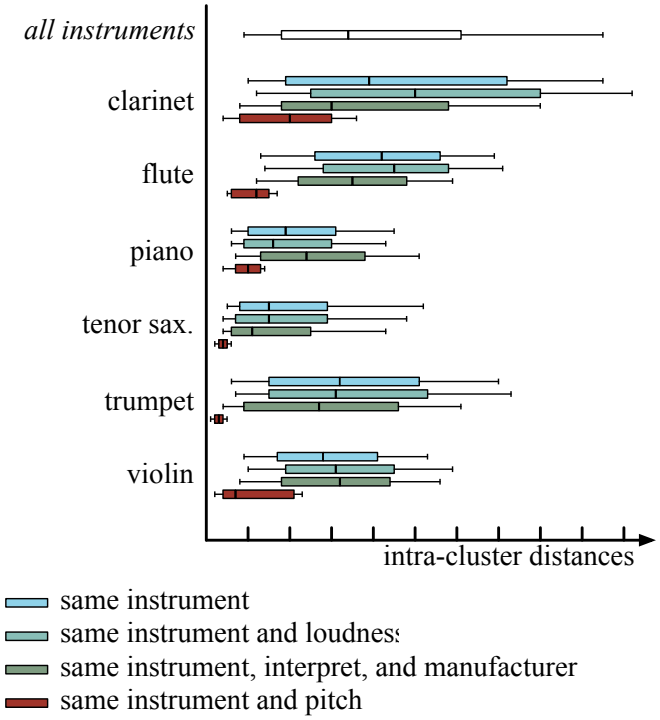


Figure 2.7: Invariants of the mel-frequency cepstrum. Distributions of squared Euclidean distances among various MFCC clusters in the RWC dataset of isolated notes. Whisker ends denote lower and upper deciles.

3

TIME-FREQUENCY SCATTERING

In the previous chapter, we have introduced the scalogram $\mathbf{U_1}\boldsymbol{x}(t,\gamma)$ of an audio signal $\boldsymbol{x}(t)$, a nonnegative matrix indexed by time t and log-frequency γ. We have shown that a well-designed auditory filterbank offers an accurate time-frequency localization of nonstationary components in the scalogram, such as sharp transients and chirps. However, the scalogram only demodulates auditory information up to a typical time scale of 20 ms. A desirable representation for audio classification should integrate a broader context while segregating independent sources.

To address this problem, this chapter introduces a new invariant representation for audio classification, named time-frequency scattering (TFS). The core idea behind a scattering transform is to convolve the scalogram with a modulation filter bank of wavelets, and subsequently apply a complex modulus nonlinearity.

In the earliest definition of the scattering transform, convolutions are performed solely upon the time variable t (Andén and Mallat, 2014). The main contribution of this chapter is to extend the temporal scattering transform to encompass both the time variable t and the log-frequency variable γ, thus improving its discriminative power. The resulting representation, called time-frequency scattering transform, corresponds to the spectrotemporal receptive fields (STRF) of Chi, Ru, and Shamma (2005), an idealized computational model for the response of neurons in the central auditory system.

The rest of this chapter is organized as follows. In Section 3.1, we discuss the limitations of the temporal scattering transform. In Section 3.2, we explain how time-frequency scattering mitigates these limitations at the expense of an increased number of coefficients. In Section 3.3, we re-synthesize locally stationary audio textures from scattering coefficients, and present a creative application in partnership with composer Florian Hecker. In Section 3.4, we report state-of-the-art results in short urban sound classification and competitive results in binaural scene classification.

Part of the content of this chapter has been previously published by Andén, Lostanlen, and Mallat (2015).

3.1 TEMPORAL SCATTERING

Amplitude modulation features (Alam et al., 2013; Gillet and Richard, 2004; Mitra et al., 2012), also called dynamic features (Peeters, La Burthe, and Rodet, 2002), have long been part of the typical set of audio

descriptors for automatic classification. By decomposing short-term spectra into a redundant filter bank of temporal modulations, they measure the intermittency of auditory information. In music, amplitude modulations are typical of note onsets as well as extended instrumental techniques. Among acoustic scenes, they may be caused by a broad variety of mechanical interactions, including collision, friction, and turbulent flow. This section presents the extraction of amplitude modulation features within a wavelet framework, a theory known as deep scattering spectrum (Andén and Mallat, 2014) or temporal scattering transform.

3.1.1 *Scattering network*

A scattering network results from the composition of wavelet operators interspersed with complex modulus nonlinearities. Because the wavelet operators are convolutional, the architecture of a scattering network bears a strong resemblance with a deep convolutional network, with the important difference that the filters are hand-crafted wavelets instead of being learned from data.

In this subsection, we recall the definition of the wavelet scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$. We define the averaged scalogram $\mathbf{S}_1\boldsymbol{x}(t,\gamma)$, which is invariant to time shifts $\mathcal{T}_b$ up to a time scale T. To recover finer scales, we introduce a modulation filter bank of wavelets $\boldsymbol{\psi}_{2^{\gamma_2}}(t)$, to be applied on each subband of the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$. We define second-order wavelet modulus coefficients $\mathbf{U}_2\boldsymbol{x}(t,\gamma,\gamma)$ and second-order scattering coefficients $\mathbf{S}_2\boldsymbol{x}(t,\gamma,\gamma_2)$ after local averaging.

Scalogram and covariance to time shifts

The first layer of the temporal scattering transform yields a two-dimensional representation $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ called scalogram, indexed by time t and log-frequency $\gamma = \log_2 \omega$.

For some $b \in \mathbb{R}$, let

$$\mathcal{T}_b : \boldsymbol{x}(t) \longmapsto (\mathcal{T}_b\boldsymbol{x})(t) = \boldsymbol{x}(t+b) \tag{3.1}$$

be the translation operator. Recalling Subsection 2.2.1, the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ is covariant with time shifts, that is, it satisfies the equation

$$\mathbf{U}_1\mathcal{T}_b\boldsymbol{x}(t,\gamma) = \mathcal{T}_b\mathbf{U}_1\boldsymbol{x}(t,\gamma). \tag{3.2}$$

Time-averaged scalogram and invariance to time shifts

Features for audio classification should be made invariant, rather than covariant, to time shifts. In order to achieve invariance to time shifts $\mathcal{T}_b$ for b smaller than T, the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ is convolved with a Gaussian low-pass filter $\boldsymbol{\phi}_T(t)$ of typical duration T, that is, of

cutoff frequency $\frac{1}{T}$. This low-pass filtering yields the averaged scalogram

$$\mathbf{S}_1 \boldsymbol{x}(t, \gamma) = (\mathbf{U}_1 \boldsymbol{x} * \boldsymbol{\phi}_T)(t, \gamma), \tag{3.3}$$

which satisfies the invariance equation

$$\mathbf{S}_1(\mathcal{T}_b \boldsymbol{x})(t, \gamma) \approx \mathbf{S}_1 \boldsymbol{x}(t, \gamma) \tag{3.4}$$

as long as $b \ll T$. A downside of this operation is that the transient information in $\mathbf{U}_1 \boldsymbol{x}(t, \gamma)$ at scales finer than T are lost by this low-pass filtering, thus reducing discriminability in feature space.

Modulation filter bank of wavelets

To mitigate this issue, the temporal scattering transform recovers fine scales by convolving $\mathbf{U}_1 \boldsymbol{x}(t, \gamma)$ with wavelets whose center frequencies are above $\frac{1}{T}$ and subsequently applying complex modulus. The wavelet $\boldsymbol{\psi}_{2^{\gamma_2}}(t)$ of center frequency $\omega_2 = 2^{\gamma_2} > \frac{1}{T}$ has an expression of the form

$$\boldsymbol{\psi}_{2^{\gamma_2}}(t) = \boldsymbol{g}_{\frac{Q_2}{2^{\gamma_2}}}(t) \exp(2\pi \mathrm{i} 2^{\gamma_2} t), \tag{3.5}$$

where $\boldsymbol{g}_{\frac{Q_2}{2^{\gamma_2}}}(t)$ is a window function of duration $\frac{Q_2}{2^{\gamma_2}}$. In numerical applications, the quality factor Q_2 is set to 1 and the function $\boldsymbol{g}$ is either chosen to be a Gaussian or a Gamma function, respectively yielding Gabor or Gammatone wavelets. The latter, unlike the former, are sensitive to time reversal. We refer to Subsection 2.2.4 for a theoretical motivation of Gammatone wavelets, and to Subsection 4.5.4 for comparative results in musical instrument classification.

Scattering coefficients

Once the continuous wavelet transform has been applied to the scalogram $\mathbf{U}_1 \boldsymbol{x}(t, \gamma)$, the complex modulus nonlinearity yields

$$\mathbf{U}_2 \boldsymbol{x}(t, \gamma, \gamma_2) = |\mathbf{U}_1 \boldsymbol{x} * \boldsymbol{\psi}_{2^{\gamma_2}}|(t, \gamma), \tag{3.6}$$

a three-dimensional tensor indexed by time t, acoustic log-frequency γ, and modulation log-frequency γ_2.

The last operation in the computation of the temporal scattering transform is the convolution between the tensor $\mathbf{U}_2 \boldsymbol{x}(t, \gamma, \gamma_2)$ and the low-pass filter $\boldsymbol{\phi}_T(t)$, yielding

$$\begin{aligned} \mathbf{S}_2 \boldsymbol{x}(t, \gamma, \gamma_2) &= (\mathbf{U}_2 \boldsymbol{x} \overset{t}{*} \boldsymbol{\phi}_T)(t, \gamma, \gamma_2) && (3.7) \\ &= \left(|\mathbf{U}_1 \boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}}| * \boldsymbol{\phi}_T \right)(t). && (3.8) \end{aligned}$$

Scattering coefficients then consist in the concatenation of first-order coefficients and second-order coefficients

$$\mathbf{S}\boldsymbol{x}(t, \lambda) = \begin{pmatrix} \mathbf{S}_1\boldsymbol{x}(t, \gamma) \\ \mathbf{S}_2\boldsymbol{x}(t, \gamma, \gamma_2) \end{pmatrix} \tag{3.9}$$

where the Greek literal λ denotes a generic path along first-order and second-order scattering variables. In other words, λ is either equal to a singleton of the form $\lambda = (\lambda_1) = ((\gamma))$ or equal to an ordered pair of the form $\lambda = (\lambda_1, \lambda_2) = ((\gamma), (\gamma_2))$.

A scattering network is depicted in Figure 3.1.

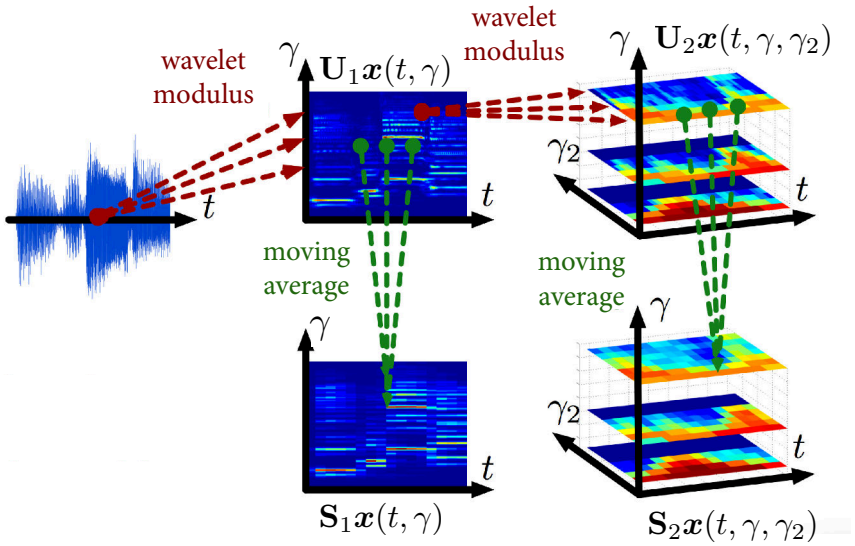


Figure 3.1: A temporal scattering network. Color shades from red to blue denote local amounts of energy from low to high.

Scattering paths

The energy of the scattering representation $\mathbf{S}\boldsymbol{x}$ is defined as

$$\|\mathbf{S}\boldsymbol{x}\|_2^2 = \left\| \begin{pmatrix} \mathbf{S}_1\boldsymbol{x} \\ \mathbf{S}_2\boldsymbol{x} \end{pmatrix} \right\|_2^2 = \|\mathbf{S}_1\boldsymbol{x}\|_2^2 + \|\mathbf{S}_2\boldsymbol{x}\|_2^2 . \tag{3.10}$$

In theory, the scattering transform is a wavelet tree of infinite depth, which integrates and demodulates transients at ever-growing temporal scales. Yet, the number of scattering coefficients increases combinatorially with depth. Moreover, Andén and Mallat (2014) have shown experimentally that, for $T = 370\,\text{ms}$, about 90% of the energy is contained in layers 1 and 2. Therefore, in most practical cases, third-order scattering coefficients

$$\mathbf{S}_3\boldsymbol{x}(t, \gamma, \gamma_2, \gamma_3) = \left(|\mathbf{U}_2\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_3}}| \overset{t}{*} \boldsymbol{\phi}_T \right)(t, \gamma, \gamma_2)$$

are neglected.

In the computation of the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$, applying the complex modulus to the first-order wavelet coefficients brings the energy of the signal from a band centered around $\omega = 2^\gamma$ to a band centered around $\omega = 0$. In the Fourier domain, the width of this band is of the order of $\frac{2^\gamma}{Q}$. Thus, scattering paths yield a negligible amount of energy if the second-order frequency 2^{γ_2} is below the first-order bandwidth $\frac{2^\gamma}{Q}$. In order to avoid unnecessary computations, we restrict $\mathbf{U}_2\boldsymbol{x}(t,\gamma,\gamma_2)$ to paths (γ,γ_2) satisfying the inequality

$$\frac{2^\gamma}{Q} > 2^{\gamma_2}, \tag{3.11}$$

also written as $\gamma > \gamma_2 + \log_2 Q$.

3.1.2 *Related work*

In the field of music information retrieval, dynamic features are historically linked to two challenges in supervised classification: musical genre recognition (Aucouturier and Pachet, 2003) and audio fingerprinting (Cano et al., 2005). The former is a coarse-grained classification problem with about ten classes and several hundred examples per class, in which intra-class variability may encompass variations in tonality, tempo, instrumentation, artist, and so forth. The latter is a fine-grained classification problem with millions of classes, in which intra-class variability is restricted to time shifts, time warps, additive noise, and distortion.

Modulation filter banks in auditory neuroscience

There is neurophysiological evidence to support the claim that the mammalian auditory system performs a temporal integration of amplitude modulations by means of a modulation filter bank (Dau, Kollmeier, and Kohlrausch, 1997; Sek and Moore, 2003). This phenomenon can be revealed by superposing two pure tones whose frequencies fall within the same critical band of the basilar membrane. We refer to Atlas and Shamma (2003) for an overview on amplitude modulation in auditory neuroscience, and to Kanedera et al. (1999) for applications to automatic speech recognition.

Musical genre recognition

Musical genre recognition relies on a loosely defined taxonomy (Sturm, 2014b). The GTZAN dataset (Tzanetakis and Cook, 2002) is the *de facto* standard in this domain. We refer to Sturm (2014b) for a recent review of the state of the art. Dynamic features in genre classification were pioneered by McKinney and Breebaart (2003), who applied a

modulation filter bank to mel-frequency cepstral coefficients (MFCC) of low dimensionality.

Since then, Andén and Mallat (2014) have shown that the concatenation of scattering transform coefficients with various quality factors Q and time scales T could lead to an unprecedented mean accuracy of 92% on the GTZAN dataset. This figure was recently relativized by Rodríguez-Algarra, Sturm, and Maruri-Aguilar (2016), who showed that the scattering pipeline was taking advantage of methodological mistakes in the construction of the GTZAN dataset. In particular, the authors have exhibited the presence of inaudible acoustic content (i.e. below 20 Hz) in some tracks, which could be an effect of the recording device instead of the genre itself. Therefore, the experiments of Andén and Mallat (2014) on GTZAN do not suffice to evaluate the efficiency of scattering representations for music content analysis.

Audio fingerprinting

The problem of audio fingerprinting, also called audio watermarking, is more well-defined than the problem of musical genre recognition. It consists in generating a different, translation-invariant hash code for every sound snippet in a music corpus and retrieve queries from radio broadcast streams (Seo et al., 2006). The use of high-dimensional dynamic features, as defined by two layers of wavelet modulus operators, was described in a patent of Rodet, Worms, and Peeters (2003).

However, current industrial solutions for audio fingerprinting, such as Shazam (Wang, 2006) or Soundhound (Mohajer et al., 2010), do not compute dynamic features. Instead, they rely on heuristics with a lower computational complexity, i.e. constructing a skeleton of ridges in the time-frequency domain, measuring the relative locations of dominant bumps, and associating a hash vector to the set of pairwise distances. Despite their massive adoption, these *ad hoc* technologies remain highly sensitive to obfuscations and are unable to retrieve cover songs. Now that on-device computational resources have improved, revisiting the original research on dynamic features could open new horizons for audio fingerprinting.

Scattering coefficients as input features for deep learning

Scattering representations can be plugged into any classification or regression system, be it shallow or deep. The original experiments of Andén and Mallat (2014) on deep scattering spectrum relied on support vector machines (SVM) with linear kernel or Gaussian kernel, as well as class-wise, affine principal component analysis (PCA). For supervised large-vocabulary continuous speech recognition, replacing these locally linear classifiers by five layers of deep neural networks (DNN) or deep convolutional networks (ConvNets) only brought marginal improvements in accuracy (Fousek, Dognin, and

Goel, 2015; Peddinti et al., 2014). However, in the Zero Resource Speech Challenge (Versteegh et al., 2015), whose aim is to discover sub-word and word units from continuous speech in an unsupervised way, associating scattering representations with deep siamese network provided a substantial gain in the tradeoff between inter-class discriminability and inter-speaker robustness (Zeghidour et al., 2016).

Scattering coefficients as input features for latent factor estimation

Aside from classification, a matrix of scattering coefficients through time can be regarded as input data for unsupervised latent factor estimation (LFE), a field of research which notably encompasses clustering and dictionary learning. Because of the modulus nonlinearity and the nonnegativity of the Gaussian low-pass filter $\boldsymbol{\phi}_T(t)$, scattering coefficients are real-valued and nonnegative by design. Therefore, a low-rank approximation of the matrix $\mathbf{S}\boldsymbol{x}$ can be obtained by nonnegative matrix factorization (Smaragdis and Brown, 2003), for applications in automatic music transcription and blind source separation (Bruna, Sprechmann, and Cun, 2015).

3.1.3 *Properties*

Four properties can be derived about the temporal scattering coefficients $\mathbf{S}_1\boldsymbol{x}(t,\gamma)$ and $\mathbf{S}_2\boldsymbol{x}(t,\gamma,\gamma_2)$. First, the scattering operator preserves the energy in the signal. Secondly, Euclidean distances between any two audio signals $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t)$ are always brought closer, never further apart, in the feature space of scattering coefficients. Thirdly, unlike the Fourier spectrum, the scattering transform is stable to small time warps $\mathcal{W}_\tau$. Fourthly, the temporal scattering transform extracts the amplitude modulation spectrum of a dynamic source-filter model, without detection or training.

Preservation of energy

The range of center frequencies for the wavelets $\boldsymbol{\psi}_{2^{\gamma_2}}(t)$ is bounded from below by the cutoff frequency $\frac{1}{T}$ of the low-pass filter $\boldsymbol{\phi}_T(t)$. Consequently, $\boldsymbol{\phi}_T(t)$ plays the role of a scaling function (Mallat, 1989) in the second-order filter bank of the temporal scattering transform. In the Fourier domain, the scaling function and the wavelets must satisfy the Littlewood-Paley inequalities

$$1-\varepsilon \leq \left|\widehat{\boldsymbol{\phi}}_T(\omega)\right|^2 + \frac{1}{2}\sum_{\gamma_2}\left(\left|\widehat{\boldsymbol{\psi}_{2^{\gamma_2}}}(\omega)\right|^2 + \left|\widehat{\boldsymbol{\psi}_{2^{\gamma_2}}}(-\omega)\right|^2\right) \leq 1 \quad (3.12)$$

for the energy in every signal to be preserved, where $\varepsilon \ll 1$ is a small nonnegative number.

By applying the Plancherel identity, Equation 3.12 rewrites as

$$1 - \varepsilon \leq \frac{\|\mathbf{S}_1 \boldsymbol{x}\|_2^2 + \|\mathbf{U}_2 \boldsymbol{x}\|_2^2}{\|\mathbf{U}_1 \boldsymbol{x}\|_2^2} \leq 1. \tag{3.13}$$

In the equation above, the constant $\varepsilon \ll 1$ ensures that the second layer of scattering transform almost preserves the energy in the scalogram $\mathbf{U}_1 \boldsymbol{x}(t, \gamma)$, which itself preserves the energy in the signal $\boldsymbol{x}(t)$. This preservation of energy suggests that discriminative information between classes is also preserved, and that the only non-injective operator in the scattering transform, namely the complex modulus nonlinearity, integrates coarse temporal context without destroying finer scales. As such, it constructs a translation-invariant representation that is not invariant to other factors of variability.

Contractiveness

Equation 3.13 implies that every linear layer of the scattering transform, understood as the combination of a wavelet filter bank and a low-pass filter $\boldsymbol{\phi}_T(t)$, is contractive. Moreover, the complex modulus satisfies the triangular inequality $||\boldsymbol{x}| - |\boldsymbol{y}|| \leq |\boldsymbol{x} - \boldsymbol{y}|$, which is also contractive. By composition of contractive operators, it follows immediately that the scattering transform is contractive, which means that two different signals $\boldsymbol{x}(t)$ belonging to the same class $\boldsymbol{y}(t)$ are only brought closer together in feature space, and not further apart. Furthermore, the shrinkage factor between distances in signal space and feature space is maximal when both signals are quasi-asymptotic and similar up to a small time shift $\mathcal{T}_b$, where $b \ll T$.

Stability to small time warps

By resorting to wavelet transform modulus, as opposed to Fourier transform modulus, scattering features are provably stable to small time warps, in the sense of Lipschitz regularity with respect to diffeomorphisms

$$\mathcal{W}_\tau : \boldsymbol{x}(t) \longmapsto (\mathcal{W}_\tau \boldsymbol{x})(t) = \dot{\boldsymbol{\tau}}(t)(\boldsymbol{x} \circ \tau)(t)$$

whose maximum amount of deformation, as measured by the supremum norm $\|\dot{\boldsymbol{\tau}}\|_\infty$ of the $\boldsymbol{\tau}(t)$, is bounded from above by $2^{1/Q}$, and $Q = 24$ is the quality factor of the auditory filter bank. The choice of Q may therefore be understood as a tradeoff between stability to time warps and frequential localization.

The Lipschitz inequality writes as

$$\|\mathbf{S}\boldsymbol{x} - \mathbf{S}\mathcal{W}_\tau \boldsymbol{x}\| \leq C \times \|\boldsymbol{x}\|_2 \times \|\dot{\boldsymbol{\tau}}\|_\infty, \tag{3.14}$$

where C is a small constant. The proof of the stability of the scattering transform to small diffeomorphisms was achieved by Mallat

(2012). This stability result implies that the wavelet modulus operator $\mathbf{U}_1$ brings warped examples $\{\mathcal{W}_\tau \boldsymbol{x}(t)\}_\tau$ to an affine, low-dimensional subspace of $L^2(\mathbb{R})$ in the vicinity of $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$. As a consequence, independent sources of variability, e.g. operating at independent time scales, are disentangled and linearized in the scattering representation $\mathbf{S}\boldsymbol{x}(t,\lambda)$. Because the scattering transform linearizes geometric intra-class variability, a linear classifier trained on scattering coefficients can outperform a nonlinear classifier trained on short-term audio descriptors (Andén and Mallat, 2014).

Extraction of amplitude modulation spectrum

The stationary source-filter model of sound production is built by convolving a harmonic excitation $\boldsymbol{e}(t) = \sum_{p=1}^{P} \cos(2\pi p \xi t)$ of fundamental frequency ξ with a broadband filter $\boldsymbol{h}(t)$. Let $\boldsymbol{\alpha}(t) > 0$ an amplitude function whose rate of variation is slow in front of ξ. The amplitude-modulated model is defined as

$$\boldsymbol{x}(t) = \boldsymbol{\alpha}(t) \times (\boldsymbol{e} * \boldsymbol{h})(t). \tag{3.15}$$

According to Andén and Mallat (2012), the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ of $\boldsymbol{x}(t)$ obtained by constant-Q wavelets such that the quality factor Q is greater than the number P of partials in the harmonic excitation is approximately equal to

$$\mathbf{U}_1\boldsymbol{x}(t,\gamma) \approx \boldsymbol{\alpha}(t) \times |\hat{\boldsymbol{h}}(2^\gamma)| \times \mathbf{U}_1\boldsymbol{e}(t,\gamma).$$

It follows that the ratio between second-order scattering coefficients $\mathbf{S}_2\boldsymbol{x}(t,\gamma,\gamma_2)$ and the corresponding first-order coefficients $\mathbf{S}_1\boldsymbol{x}(t,\gamma)$ yields nonparametric measurements of the amplitude modulation spectrum

$$\frac{\mathbf{S}_2\boldsymbol{x}(t,\gamma,\gamma_2)}{\mathbf{S}_1\boldsymbol{x}(t,\gamma)} \approx \frac{\left(|\boldsymbol{\alpha} * \boldsymbol{\psi}_{2^{\gamma_2}}| * \boldsymbol{\phi}_T\right)(t)}{(\boldsymbol{\alpha} * \boldsymbol{\phi}_T)(t)}. \tag{3.16}$$

It should be observed that the coefficients above are not only invariant to time shifts, but also to frequency transposition, and to any modification of $\boldsymbol{h}(t)$.

Besides the amplitude modulation spectrum, we refer to Subsection 2.1.3 and to the article of Andén and Mallat (2014) for insights on how the temporal scattering transform can measure frequency intervals from interferences.

3.1.4 *Limitations*

In this subsection, we pinpoint the limitations of the temporal scattering transform for audio classification. We construct a frequency-dependent translation operator in the time-frequency domain, and

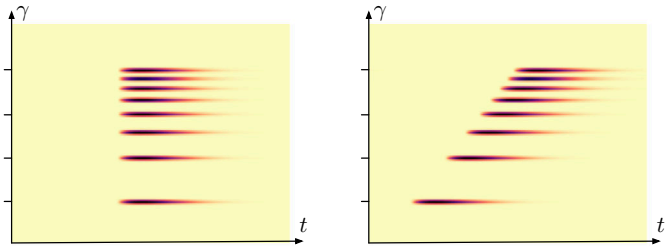


Figure 3.2: The action of frequency-dependent time shift on an additive model of $P = 8$ partials. In this example, the time shift is an affine function of log-frequency. Observe that onsets are no longer synchronized after the application of the time shift.

show that the temporal scattering transform is insensitive to the action of this operator. Moreover, we argue that the result of Andén and Mallat (2014) on the source-filter model relies on stationarity assumptions for both the source and the filter, thus restricting its time scale of validity.

Insensitivity to frequency-dependent time shifts

The low-pass filtering with $\boldsymbol{\phi}_T(t)$ while going from $\mathbf{U}\boldsymbol{x}(t,\lambda)$ to $\mathbf{S}\boldsymbol{x}(t,\lambda)$ guarantees that scattering coefficients in $\mathbf{S}\boldsymbol{x}(t,\gamma)$ are invariant to time shifts $\mathcal{T}_b$ for any $b \ll T$. However, the temporal scattering transform is also invariant to a larger group of transformations, called frequency-dependent time shifts.

A frequency-dependent time shift is built by defining a regular, yet non-constant function $\boldsymbol{b}(\gamma)$, and applying the translation $\mathcal{T}_{\boldsymbol{b}(\gamma)}$ to every subband γ in the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$. When applied to an additive sinusoidal model with a sharp attack at $t = 0$, the action of $\mathcal{T}_{\boldsymbol{b}(\gamma)}$ misaligns the onset times of neighboring partials, as it brings them further apart from zero. The resulting misalignment is depicted in Figure 3.2. Although ill-defined in the time domain, it boils down to the geometrical transformation

$$(\mathcal{T}_{\boldsymbol{b}(\gamma)}\mathbf{U}_1\boldsymbol{x})(t,\gamma) = \mathbf{U}_1\boldsymbol{x}(t+\boldsymbol{b}(\gamma),\gamma) \tag{3.17}$$

in the time-frequency domain.

We denote by

$$(\mathbf{S}_2\mathcal{T}_{\boldsymbol{b}(\gamma)}\boldsymbol{x})(t,\gamma,\gamma_2) = \left(\left|(\mathcal{T}_{\boldsymbol{b}(\gamma)}\mathbf{U}_1\boldsymbol{x}) * \boldsymbol{\psi}_{2^{\gamma_2}}\right| * \boldsymbol{\phi}\right)(t,\gamma)$$

the second-order scattering coefficients of the deformed scalogram. Because the modulation filter bank is a convolutional operator, it commutes with the time shift $\mathcal{T}_{\boldsymbol{b}(\gamma)}$ for every log-frequency γ:

$$\mathbf{U}_2\mathcal{T}_{\boldsymbol{b}(\gamma)}\boldsymbol{x}(t,\gamma,\gamma_2) \quad = \quad \mathbf{U}_2\boldsymbol{x}(t+\boldsymbol{b}(\gamma),\gamma,\gamma_2).$$

Assuming that the function $\boldsymbol{b}$ takes values between $-\frac{T}{2}$ and $\frac{T}{2}$, the application of the low-pass filter $\boldsymbol{\phi}_T(t)$ while going from $\mathbf{U}_2\boldsymbol{x}$ to $\mathbf{S}_2\boldsymbol{x}$

blurs out the effect of the frequency-dependent translation $\mathcal{T}_{\boldsymbol{b}(\gamma)}$. Consequently, the scattering representation $\mathbf{S}\mathcal{T}_{\boldsymbol{b}(\gamma)}\boldsymbol{x}(t,\lambda)$ of the deformed signal is indistinguishable from the scattering representation of the original signal $\mathbf{S}\boldsymbol{x}(t,\lambda)$. The resulting inequality

$$\|\mathbf{S}\boldsymbol{x}(t,\lambda) - (\mathbf{S}\mathcal{T}_{\boldsymbol{b}(\gamma)}\boldsymbol{x})(t,\lambda)\| \ll \|\mathbf{S}\boldsymbol{x}(t,\lambda)\|$$

will be experimentally validated in Subsection 3.2.3.

Insensitivity to joint time-frequency patterns

In the context of the source-filter model, a limitation of the temporal scattering transform is that slow modulations cannot be retrieved from $\mathbf{S}\boldsymbol{x}(t,\lambda)$ if either the source or the filter themselves have varying properties over time. In speech or music, such hypotheses are only valid up to a typical time scale of $T = 25$ ms, but become too simplistic at larger time scales. Subsection 4.3.1 addresses a generalized definition of the source-filter model, in which $\boldsymbol{\alpha}(t)$, $\boldsymbol{e}(t)$, and $\boldsymbol{h}(t)$ are jointly deformed over time. It will be shown that time-frequency scattering and spiral scattering are more appropriate tools than temporal scattering to address the generalized case.

3.2 TIME-FREQUENCY SCATTERING

The temporal scattering transform presented in the previous section provides a representation of sounds which is invariant to time shifts and stable to time warps. However, because it scatters scalogram subbands independently, it has the undesirable property of being invariant to a larger group of time shifts, called frequency-dependent time shifts. As a consequence, the spectrotemporal coherence is lost after application of the complex modulus and of the low-pass filter.

In this section, we address the problem of designing an improved version of the temporal scattering transform which characterizes time-frequency patterns. To do so, we replace the second layer of the temporal scattering transform by the composition of two modulation filter banks, one over time and one over log-frequency. The resulting operator, named time-frequency scattering, extracts edges in the time-frequency domain, which are typical of chirps and broadband impulses. Time-frequency scattering is similar to the output magnitude of the "cortical representation" of Chi, Ru, and Shamma (2005), an idealized model of neural responses in the central auditory system

First, we define time-frequency scattering as a composition of convolutional operators and complex modulus nonlinearities. Secondly, we review similar ideas to time-frequency scattering in the scientific literature. Thirdly, we show two properties of time-frequency scattering which are not available to temporal scattering, namely sensitivity to frequency-dependent time shifts and extraction of chirp rate.

3.2.1 *Spectrotemporal filter bank*

In this subsection, we address the problem of composing two wavelet transforms over the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$, one along the time variable t and one along the log-frequency variable γ. To begin with, we provide a formalism to represent the semantics of the variables produced by wavelet transforms, such as scale and orientation. Then, we define the two modulation filter banks associated with the variables t and γ. Lastly, we multiply wavelets over t and γ to obtain a two-dimensional filter bank in the time-frequency domain.

Signed frequencies

Three variables come into play in the wavelet transform of a one-dimensional, complex-valued signal: a spatial or temporal variable t, a log-frequency variable γ, and an orientation variable θ. Both t and γ take values in the real line $\mathbb{R}$, whereas θ takes values in the set $\{-1, 0, 1\}$, respectively denoting negative, null, and positive frequencies. The center frequency of the wavelet $\boldsymbol{\psi}_\omega(t)$ writes as

$$\omega = \theta \times 2^\gamma,$$

where the log-frequency index $\gamma = \log_2 |\omega|$ takes evenly spaced values. In the edge case of a null center frequency $\omega = 0$, the wavelet is replaced by a low-pass filter $\boldsymbol{\phi}_T(t)$, the orientation variable is set to zero and the log-frequency variable is set to $\gamma = -\infty$ by convention.

In the case of audio signals, the input $\boldsymbol{x}(t)$ is real-valued instead of complex-valued. Consequently, its Fourier transform is Hermitian symmetric:

$$\hat{\boldsymbol{x}}(-\omega) = \hat{\boldsymbol{x}}^*(\omega).$$

As a result, one may restrict the constant-Q filter bank to positive frequencies, that is, discard the values $\theta = -1$ and $\theta = 0$. The set of values taken by the orientation variable θ boils down to a singleton, and the scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$ is a function of merely two variables instead of three.

Likewise, the modulation filter bank of the temporal scattering transform is applied over the scalogram, which is real-valued. Thus, the orientation variable θ_2 is not needed either in the modulation filter bank along time, whose center frequencies are of the form $\omega = 2^{\gamma_2}$. Second-order wavelet modulus coefficients write as a real-valued, three-way tensor

$$\mathbf{U}_2\boldsymbol{x}(t,\gamma,\gamma_2) = |\mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}}|(t,\gamma).$$

The situation is different for time-frequency scattering. Indeed, the three-dimensional tensor $\mathbf{Y}_2\boldsymbol{x}(t,\gamma,\gamma_2) = (\mathbf{U}_1\boldsymbol{x} * \boldsymbol{\psi}_{2^{\gamma_2}})(t,\gamma)$ is complex-valued, not real-valued. Therefore, the Fourier transform of $\mathbf{Y}_2\boldsymbol{x}(t,\gamma,\gamma_2)$

computed over the pseudo-continuous variable γ is no longer Hermitian symmetric. In order to preserve information in $\mathbf{Y}_2 \boldsymbol{x}(t, \gamma, \gamma)$, the modulation filter bank over log-frequencies should encompass negative frequencies as well as positive frequencies. Therefore, time-frequency scattering yields a tensor of five variables: t, γ, γ_2, log-frequency along γ, and orientation along γ.

Variables as lists

To express these five variables, and in particular the latter two, we choose to adopt a notation which could convey the semantics of the word "along". We propose to view every variable in a scattering network as a list whose head is either of the form γ_m (log-frequency) or θ_m (orientation), and whose tail is the variable along which the wavelet transform has been performed. If the depth m is not explicit, it default to 1.

In order to construct a list L whose head is the symbol h and whose tail is T, we write

$$L = h::T.$$

The operator ::, called list construction and abbreviated as *cons*, is borrowed from the ML family of programming languages (Milner, 1978). It is right-associative and has a stronger precedence than the usual arithmetical operators. Because all lists terminate with the symbol t, we elide this symbol so as to alleviate notation.

The five variables involved in time-frequency scattering are

1. time t;
2. acoustic log-frequency $\gamma_1::t$, elided to γ;
3. second-order log-frequency along time $\gamma_2::t$, elided to γ_2;
4. second-order log-frequency along log-frequency $\gamma_1::\gamma_1::t$, elided to $\gamma::\gamma$; and
5. second-order orientation along log-frequency $\theta_1::\gamma_1::t$, elided to $\theta::\gamma$.

The formalism which assigns a list of literals to every variable in the scattering network is implemented in our implementation of the time-frequency scattering transform, whose source code is available at the address `github.com/lostanlen/scattering.m`, and released under an MIT license.

Modulation filter bank along log-frequencies

Let $\boldsymbol{\psi}(\gamma)$ be a Gabor wavelet of dimensionless center frequency 1. A modulation filter bank along the variable γ is built by dilating $\boldsymbol{\psi}(\gamma)$ by scales $2^{-\gamma::\gamma}$, thus yielding wavelets of the form

$$\boldsymbol{\psi}_{\theta::\gamma \times 2^{\gamma::\gamma}}(\gamma) = \begin{cases} 2^{\gamma::\gamma}\boldsymbol{\psi}((\theta::\gamma) \times 2^{\gamma::\gamma}\gamma) & \text{if } (\theta::\gamma) \neq 0, \\ \boldsymbol{\phi}_{T::\gamma}(\gamma) & \text{otherwise} \end{cases}.$$

The purpose of the orientation variable $\theta::\gamma$ is to flip the center frequency of the wavelet from $2^{\gamma::\gamma}$ to $-2^{\gamma::\gamma}$, thus reaching negative frequencies. As for the null frequency $\omega::\gamma = 0$, it is covered by the edge case $\theta::\gamma = 0$, where the wavelet is replaced by a low-pass filter $\boldsymbol{\phi}_{T::\gamma}(\gamma)$ of size $T::\gamma$. In numerical experiments, $T::\gamma$ is chosen to be of the order of four octaves.

The center frequency along γ is thus equal to

$$\omega::\gamma = \theta::\gamma \times 2^{\gamma::\gamma}.$$

As seen previously in Subsection 3.1.3, the modulation filter bank along log-frequencies γ is designed to satisfy the Littlewood-Paley inequalities (Equation 3.12), which implies that the wavelet operator preserves energy, contracts distances, and remains stable to small deformations.

Two-dimensional filter bank

Before applying the temporal averaging $\boldsymbol{\phi}(t)$, second-order temporal scattering coefficients $\mathbf{U}_2\boldsymbol{x}(t, \gamma)$ are equal to

$$\mathbf{U}_2\boldsymbol{x}(t, \gamma, \gamma_2) = \left| \mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}} \right| (t, \gamma). \tag{3.18}$$

In comparison, second-order time-frequency scattering coefficients are defined as

$$\mathbf{U}_2\boldsymbol{x}(t, \gamma, \gamma_2, \gamma::\gamma, \theta::\gamma) = \left| \mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}} \overset{\gamma}{*} \boldsymbol{\psi}_{(\theta::\gamma)\times 2^{\gamma::\gamma}} \right| (t, \gamma). \tag{3.19}$$

Because they operate over distinct variables, the convolutions along t and along γ can be factorized into one two-dimensional convolution in the time-frequency domain (t, γ) with a wavelet

$$\boldsymbol{\Psi}_{\lambda_2}(t, \gamma) = \boldsymbol{\psi}_{2^{\gamma_2}}(t) \times \boldsymbol{\psi}_{(\theta::\gamma)\times 2^{\gamma::\gamma}}(\gamma),$$

where the generic index $\lambda_2 = (\gamma_2, \gamma::\gamma, \theta::\gamma)$ encompasses all three log-frequency and orientation variables involved in the computation of second-order scattering coefficients. The definition of $\mathbf{U}_2\boldsymbol{x}$ in Equation 3.19 can be rewritten in short as

$$\mathbf{U}_2\boldsymbol{x}(t, \gamma, \lambda_2) = |\mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\Psi}_{\lambda_2}|(t, \gamma),$$

similarly to 3.18.

Three joint time-frequency wavelets $\boldsymbol{\Psi}_{\lambda_2}(t, \gamma)$ for various values of the multi-index variable $\lambda_2 = (\gamma_2, \gamma::\gamma, \theta::\gamma)$ are shown in Figure 3.3.

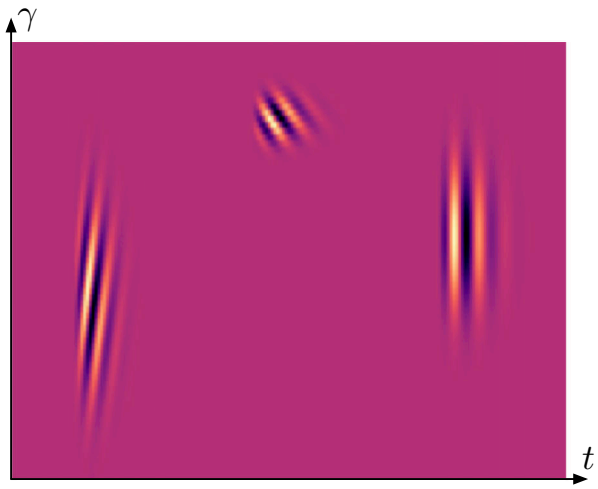


Figure 3.3: Joint time-frequency wavelets in the time-frequency domain. Brighter colors denote greater algebraic values of the real part. The background color corresponds to a null real part.

3.2.2 *Related work*

The principal motivation behind time-frequency scattering is to capture the regularity of scalogram ridges while segregating independent spectrotemporal scales. These two principles of integration and segregation are firmly established in the field of acoustic scene analysis. Yet, they are also found in signal processing and auditory neuroscience.

In this subsection, we review some previous work showing that the extraction of local contrast features in the time-frequency domain provides a sparse, meaningful representation of auditory information. Within the field of signal processing, we summarize the state of the art in time-frequency reassignment, chirplets, matching pursuit with time-frequency atoms. Within the field of auditory neuroscience, we present the spectrotemporal receptive fields (STRF), which are closely similar to time-frequency scattering. Lastly, we review the use of spectrotemporal features for audio classification, and describe the (separable) time and frequency scattering transform as a predecessor of the (joint) time-frequency scattering transform.

Time-frequency reassignment

In the field of signal processing, the problem of capturing spectrotemporal modulations in complex sounds was first addressed by means of the Wigner-Ville distribution

$$\mathrm{WVD}(\boldsymbol{x})(t,\omega) = \int \boldsymbol{x}\left(t+\frac{t'}{2}\right)\boldsymbol{x}^*\left(t-\frac{t'}{2}\right)\exp(-2\pi\mathrm{i}\omega t')\,\mathrm{d}t',$$

a quadratic function of two variables, time t and frequency ω (Flandrin, 1998). The Wigner-Ville distribution (WVD) is agnostic to the Heisenberg tradeoff in time-frequency localization induced by the choice of a convolutional operator, such as a short-term Fourier transform (STFT), a continuous wavelet transform (CWT), or a nonstationary Gabor transform (NSGT). Therefore, the WVD has a perfect time-frequency localization of sine waves, Diracs, as well as all linear chirps of the form

$$x(t) = \cos\left(\varphi_0 + \varphi_1 t + \frac{\varphi_2}{2} t^2\right).$$

On the flip side, the WVD of a signal containing multiple asymptotic components presents interferences between components, which hinders its readability.

In order to improve the time-frequency localization of classical time-frequency representations while avoiding interferences, a family of adaptive algorithms in time-frequency reassignment has been developed. We refer to Flandrin, Auger, Chassande-Mottin, et al. (2002) for an overview.

Chirplets and matching pursuit

A parallel branch of research in signal processing has been devoted to the construction of a linear decomposition of audio signals that would capture frequency modulation as well as amplitude modulation. The invention of the chirplet, i.e. a linear chirp with a Gabor envelope, is particularly relevant in that regard (Baraniuk and Jones, 1992). The chirplet transform of Mann and Haykin (1995) maps every signal to a three-dimensional domain indexed by time, frequency, and chirp rate. In turn, the chirplet representation of Bultan (1999) is indexed by four variables: time, frequency, temporal scale, and angle in the time-frequency plane.

There are two major differences between the chirplet transform and the time-frequency scattering transform. First, frequencies in the chirplet transform are linearly spaced instead of geometrically spaced. Secondly, the chirplet transform operates directly on the waveform instead of the wavelet scalogram $\mathbf{U}_1 x(t, \gamma)$. As a consequence, the chirplet transform is a linear operator with respect to the signal $x(t)$, whereas the time-frequency scattering transform is nonlinear because of the application of complex modulus.

Owing to their good time-frequency localization, wavelets provide a sparse representation of speech and music. Representing these signals with few time-frequency atoms is at the heart of spare coding algorithms, such as matching pursuit and its variants (Mallat and Zhang, 1993). Because of the abundance of chirps in natural signals, replacing wavelets by a redundant dictionary of Gaussian chirps enables an even greater sparsity (Gribonval, 2001).

Spectrotemporal receptive fields

In the field of neurophysiology, there is empirical evidence supporting the claim that the auditory system extracts spectrotemporal modulations at various scales and rates. To demonstrate it, scientists insert electrodes into the primary auditory cortex (A1) of a living animal, thus measuring spike trains emitted by isolated neurons. Then, the animal is exposed to a random sequence of stimuli. The spike train elicited by each stimulus is integrated through time with a low-pass filter, yielding the post-stimulus time histogram (PSTH). Finally, the spectrotemporal receptive field (STRF) of the observed neuron is computed as the optimal predictor of its PSTH.

The simplest way to obtain the STRF of a neuron is the reverse correlation method (De Boer and Kuyper, 1968), which consists in averaging all stimuli that trigger a spike. The STRF is then proportional to the spike-triggered average stimulus (STA) if the following three conditions are met:

1. the stimulus space encompasses all stimuli that are capable of eliciting spikes,
2. the sampling in stimulus space is random and uniform, and
3. the multiple spatial dimensions are independent.

Although the study of the visual cortex (V1) benefits from a canonical representation of images as bidimensional arrays of pixels, the study of STRFs in audition is conditional upon the definition of a time-frequency representation, in which the frequential dimension plays the same role as the two spatial dimensions in vision.

Initially, the STRF was defined as the first-order Volterra kernel that relates the Wigner-Ville distribution of the sound pressure waveform to the PSTH (Aertsen and Johannesma, 1981; Eggermont, 1993). This definition is agnostic to the choice of a Heisenberg tradeoff in time-frequency localization, but, again, its readability is hindered by the presence of interferences. A concurrent definition, the "spectrographic STRF" (Klein et al., 2000; Theunissen, Sen, and Doupe, 2000), relies on a smoothing of the Wigner-Ville STRF by a convolutional time-frequency kernel.

The Gammatone scalogram introduced in Subsection 2.2.4 is a valid computational model for the frequential selectivity of hair cells in the basilar membrane of the cochlea. Therefore, it makes sense to define the spectrographic STRF as operating over the Gammatone scalogram. Yet, because wavelets in the auditory filter bank overlap in the Fourier domain, the scalogram $\mathbf{U}_1 \boldsymbol{x}(t, \gamma)$ is a redundant representation of the waveform $\boldsymbol{x}(t)$. Consequently, we must recall that the spike-triggered average stimulus (STA) is only an approximation of the true STRF.

The STRF based on the STA does not only depend upon the choice of time-frequency representation, but also on the statistical properties

of the stimuli. Despite the fact that white noise satisfies the condition 2 above, it does not conform with condition 1, as it only elicits few spikes per neuron. In contrast, more salient patterns, such as dynamic ripples, tend to elicit sparser responses in the neural domain (Depireux et al., 2001).

Spectrotemporal features for classification

In the field of pattern recognition, the previous decade has seen the emergence of idealized models of STRF, to be used as features for classification (Chi, Ru, and Shamma, 2005). We refer to Lindeberg and Friberg (2015) for a recent overview of such idealized models.

At first, they have been used in automatic speech recognition (Kleinschmidt, 2002), as well as binary classification of speech versus non-speech (Mesgarani, Slaney, and Shamma, 2006). More recently, they have been regarded as general-purpose timbral features (Siedenburg, Fujinaga, and McAdams, 2016), and used for musical instrument classification (Patil and Elhilali, 2015; Patil et al., 2012) and speaker recognition (Lei, Meyer, and Mirghafori, 2012). Under the name of Gabor filter bank features (Schädler, Meyer, and Kollmeier, 2012), they have led to state-of-the-art results in acoustic event detection (Schröder et al., 2013).

The idealized STRF model, sometimes denoted as full cortical model (Patil et al., 2012), is extracted from the averaged scalogram $\mathbf{S}_1\boldsymbol{x}(t,\gamma)$ as

$$\mathbf{U}_2\boldsymbol{x}(t,\gamma,\gamma_2,\gamma{::}\gamma,\theta{::}\gamma) = \left|\mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}} \overset{\gamma}{*} \boldsymbol{\psi}_{(\theta::\gamma)\times 2^{\gamma::\gamma}}\right| (t,\gamma).$$

Within a discrete framework, the critical sample rate between adjacent log-frequencies γ in $\mathbf{U}_2\boldsymbol{x}$ is proportional to $2^{\gamma::\gamma}$. However, the full cortical model does not subsample adjacent log-frequencies, as it returns a dense tensor of coefficients. As a result, the full cortical model has more coefficients than time-frequency scattering, which is sampled at critical rates. We refer to Hemery and Aucouturier (2015) for an overview of dimensionality reduction techniques of the full cortical model in the context of environmental sound classification.

Separable time and frequency scattering

One of the major questions raised by the analysis of spectrotemporal receptive fields is whether or not they are separable along time and frequency. Qiu, Schreiner, and Escabí (2003) have used singular value decomposition (SVD) to approximate STRFs by a sum of products of real-valued Gabor filters in time and frequency. They found that 60% of neurons in the inferior colliculus (IC) of the cat are well described by a separable product of a real-valued Gabor in time and a real-

valued Gabor in frequency. The behavior of these neurons can be put in parallel with the separable model of time and frequency scattering:

$$\mathbf{U}_2\boldsymbol{x}(t,\gamma,\gamma_2,\gamma::\gamma) = \left| |\mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}}| \overset{t}{*} \boldsymbol{\phi}_T \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma::\gamma}} \right| (t,\gamma),$$

in which the application of modulation filter banks along time and log-frequency are interspersed by a complex modulus nonlinearity (Andén and Mallat, 2014).

Conversely, the remaining 40% of neurons exhibit obliquely oriented receptive fields, along with lateral subfields of excitation and inhibition. The STRFs of these remaining neurons are out of the scope of separable time and frequency scattering, but can be fitted by joint time-frequency scattering

$$\mathbf{U}_2\boldsymbol{x}(t,\gamma,\lambda_2) = \left| \mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\Psi}_{\lambda_2} \right| (t,\gamma),$$

where λ_2 is a multi-index denoting the shape of the convolutional operator $\boldsymbol{\Psi}_{\lambda_2}$ in the time-frequency domain.

We refer to Schädler and Kollmeier (2015) for a comparison of separable and nonseparable models of Gabor filter bank (GBFB) features in the context of environmental sound classification.

3.2.3 *Properties*

In this subsection, we describe two properties satisfied by time-frequency scattering that are not satisfied by temporal scattering. First, time-frequency scattering is sensitive to frequency-dependent time shifts. Secondly, time-frequency scattering is capable of extracting the local chirp rate of an asymptotic signal, and distinguish upward chirps from downward chirps.

Sensitivity to frequency-dependent time shifts

In Subsection, 3.1.4, we have seen that the temporal scattering transform is insensitive to frequency-dependent time shifts $\mathcal{T}_{\boldsymbol{b}(\gamma)}$, where $\boldsymbol{b}(\gamma)$ is a regular but non-constant function of log-frequency γ. Here, we show that time-frequency scattering is affected by such frequency-dependent time shifts.

Assuming that the inequality

$$\left| \frac{\partial \boldsymbol{b}}{\partial \gamma} \right| (\gamma') \ll 2^{-\gamma'} Q$$

holds for every log-frequency γ', the variations of the function $\boldsymbol{b}(\gamma)$ are small enough to be locally neglected over the log-frequential range of every auditory filter. As a consequence, $\mathcal{T}_{\boldsymbol{b}(\gamma)}$ is well defined as

$$\mathcal{T}_{\boldsymbol{b}(\gamma)}\mathbf{U}_1\boldsymbol{x}(t,\gamma) \approx \mathbf{U}_1\boldsymbol{x}(t+\boldsymbol{b}(\gamma),\gamma).$$

Because the scattering transform is a convolutional operator, it commutes with a well-defined $\mathcal{T}_{\boldsymbol{b}(\gamma)}$:

$$\begin{aligned}\mathbf{Y}_2\mathcal{T}_{\boldsymbol{b}(\gamma)}\boldsymbol{x}(t,\gamma,\gamma_2) &= \mathcal{T}_{\boldsymbol{b}(\gamma)}(\mathbf{U}_1\boldsymbol{x})(t,\gamma) \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}}(t) \\ &\approx \mathcal{T}_{\boldsymbol{b}(\gamma)}(\mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}})(t,\gamma) \\ &= \mathcal{T}_{\boldsymbol{b}(\gamma)}\mathbf{Y}_2\boldsymbol{x}(t,\gamma,\gamma_2).\end{aligned}$$

Consequently, the averaged temporal scattering coefficients

$$\mathbf{S}_2\boldsymbol{x}(t,\gamma,\gamma_2) = \left(|\mathbf{Y}_2\boldsymbol{x}| \overset{t}{*} \boldsymbol{\phi}_T\right)(t,\gamma,\gamma_2)$$

are not sensitive to the action of $\mathcal{T}_{\boldsymbol{b}(\gamma)}$ as long as the range of values taken by $\boldsymbol{b}(\gamma)$ does not exceed T.

The situation is different with time-frequency scattering. Indeed, adjacent subbands are transformed with a modulation filter bank along log-frequencies γ, which constrains the relative locations of temporal cues, such as onsets. For a log-frequential scale $2^{-\gamma::\gamma}$ large enough so that the variations of $\boldsymbol{b}(\gamma)$ are noticeable at the temporal modulation scale of $2^{-\gamma_2}$, the action of the frequency-dependent time shift affects time-frequency scattering coefficients:

$$\begin{aligned}\mathbf{Y}_2\mathcal{T}_{\boldsymbol{b}(\gamma)}\boldsymbol{x}(t,\gamma,\lambda_2) &= \mathcal{T}_{\boldsymbol{b}(\gamma)}(\mathbf{U}_1\boldsymbol{x})(t,\gamma) \overset{t,\gamma}{*} \boldsymbol{\Psi}_{\lambda_2}(t,\gamma) \\ &\neq \mathcal{T}_{\boldsymbol{b}(\gamma)}\mathbf{Y}_2\boldsymbol{x}(t,\gamma,\lambda_2).\end{aligned}$$

In order to validate this fact experimentally, we compute the temporal scattering transform and time-frequency scattering transform of an additive sinusoidal model of eight partials, before and after frequency-dependent time shift. The scalogram of this sinusoidal model is shown in Figure 3.2. Then, we compute the ratios

$$\frac{\|\mathbf{S}_2\boldsymbol{x} - \mathbf{S}_2\mathcal{T}_{\boldsymbol{b}(\gamma)}\boldsymbol{x}\|_2}{\|\mathbf{S}_2\boldsymbol{x}\|_2}$$

for $\mathbf{S}_2$ being a temporal scattering operator and a time-frequency scattering operator. We obtain a ratio of 1.6×10^{-4} for temporal scattering and a ratio of 0.44 for time-frequency scattering. The gap between these two ratios confirms the theory.

Extraction of chirp rate

Another important property of time-frequency scattering is its ability to retrieve the chirp rate of a quasi-asymptotic signal $\boldsymbol{x}(t) = \cos(2\pi\boldsymbol{\theta}(t))$. The wavelet ridge theorem of (Delprat et al., 1992), as explained in Subsection 2.2.3, yields an approximate formula for the scalogram

$$\mathbf{U}_1\boldsymbol{x}(t,\gamma) \approx \frac{1}{2}\widehat{g_{2^{-\gamma}}}(2^\gamma - \dot{\boldsymbol{\theta}}(t))$$

of $x(t)$. For a log-frequency $\gamma::\gamma$ small in front of the thickness of the chirp in the scalogram, we obtain

$$(\mathbf{U}_1 x \overset{\gamma}{*} \psi_{\omega::\gamma})(t,\gamma) \approx C \times \psi_{\omega::\gamma}(\gamma - \log_2 \dot{\theta}(t)), \tag{3.20}$$

where the constant C does not depend upon $x(t)$. For a fixed γ, the above is a progressive wave of instantaneous frequency equal to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(-\log_2 \dot{\theta}(t)\right) = -\frac{1}{\log 2} \times \frac{\ddot{\theta}(t)}{\dot{\theta}(t)},$$

and with a support proportional to $2^{-\gamma::\gamma}$. Convolving Equation 3.20 with a wavelet $\psi_{\omega_2}(t)$ on which this support remains approximately constant gives

$$\begin{aligned} \mathbf{Y}_2 x(t,\lambda_2) &= (\mathbf{U}_1 x \overset{t,\gamma}{*} \mathbf{\Psi}_{\lambda_2})(t,\gamma) \\ &\approx C \times \psi_{\omega::\gamma}\left(\gamma - \log_2 \dot{\theta}(t)\right) \times \widehat{\psi}_{\omega_2}\left(-\frac{\omega::\gamma}{\log 2} \times \frac{\ddot{\theta}(t)}{\dot{\theta}(t)}\right) \end{aligned}$$

The function $\widehat{\psi}_{\omega_2}$ is a smooth Gaussian bump in the Fourier domain, centered around the frequency ω_2. Consequently, the time-frequency scattering ridges of $\mathbf{U}_2 x(t,\gamma,\lambda_2) = |\mathbf{Y}_2 x|(t,\gamma,\lambda_2)$ are localized near couples $\lambda_2 = (\omega_2, \omega::\gamma)$ of frequencies satisfying the Cartesian equation

$$\omega_2 + \frac{1}{\log 2} \times \frac{\ddot{\theta}(t)}{\dot{\theta}(t)} \times (\omega::\gamma) = 0.$$

It appears from the above that the two-dimensional wavelet $\mathbf{\Psi}_{\lambda_2}(t,\gamma)$ behaves as an oriented edge extractor in the time-frequency domain, eliciting a maximal response when the orientations of the ridge and the orientation of the wavelet $\mathbf{\Psi}_{\lambda_2}(t,\gamma)$ are orthogonal. The ratio between the temporal frequency ω_2 and the log-frequential frequency $\omega::\gamma$ yields the negative chirp rate

$$\frac{\omega_2}{\omega::\gamma} = -\frac{1}{\log 2} \times \frac{\ddot{\theta}(t)}{\dot{\theta}(t)},$$

as measured in octaves per second.

3.3 AUDIO TEXTURE SYNTHESIS

Synthesizing a new signal $x(t)$ from translation-invariant coefficients $\mathbf{S}x(\lambda)$ reveals whether the auditory patterns in $x(t)$ are preserved in the representation or whether they are lost after temporal averaging. In this subsection, we present a gradient descent algorithm for signal reconstruction from scattering coefficients. We show that time-frequency scattering leads to a perceptual improvement over temporal scattering and other translation-invariant representations.

3.3.1 *Synthesis from summary statistics*

With the aim of generating realistic sound tracks of arbitrary long durations, the topic of audio texture synthesis has many applications in virtual reality and multimedia design (Schwarz, 2011). Moreover, within the field of computational neuroscience, it offers a test bed for the comparative evaluation of biologically plausible models for auditory perception (McDermott and Simoncelli, 2011).

Texture synthesis as a minimization problem

Given a target signal $x(t)$ and a translation-invariant representation $\mathbf{S}$, the problem of texture synthesis can be formulated as the minimization of the error functional

$$\begin{aligned} E(y) &= \frac{1}{2}\|\mathbf{S}y - \mathbf{S}x\|_2^2 \\ &= \frac{1}{2}\sum_{\lambda}|\mathbf{S}y(\lambda) - \mathbf{S}x(\lambda)|^2 . \end{aligned}$$

between the target representation $\mathbf{S}x(\lambda)$ and the synthesized representation $\mathbf{S}y(\lambda)$ with respect to the synthesized signal $y(t)$.

The two modulus nonlinearities in the scattering transform discard the phases of complex-valued coefficients while retaining only their magnitudes. Although the wavelet transform has a functional inverse, the knowledge of phases in every wavelet subband is needed to recover the original signal. Finding an efficient algorithm to retrieve a signal from the complex modulus of redundant linear measurements, such as convolutions with wavelets, is an open mathematical problem called phase retrieval (Mallat and Waldspurger, 2015).

In this dissertation, we address the more specific problem of resynthesizing a signal from its scattering coefficients (Bruna and Mallat, 2013a). Because of the non-convexity of the error functional E, an iterative procedure with random initialization, such as gradient descent, only converges towards a local minimum of E. However, in practice, the local minimum found by gradient descent is of relatively low error, typically 5% or less of the squared energy $\|x\|^2$ of the target. Furthermore, as will be shown, the convergence rate can be improved by the addition of a momentum term in the update, as well as the use of an adaptive learning rate policy.

State of the art

McDermott and Simoncelli (2011) have built a set of summary statistics from a representation $\mathbf{S}$ which bears a close resemblance with the temporal scattering transform, as it consists of a cascade of two constant-Q wavelet filter banks, each followed by a contractive nonlinearity. This representation has proven efficient if the target texture

$x(t)$ consists of a large number of unsynchronized, identical sources — e.g. insects, raindrops, applauding hands, and so forth. However, it fails to recover more salient perceptual attributes, such as pitch, rhythm, and reverberation.

There are two important differences between McDermott and Simoncelli's representation and the temporal scattering transform. First, a cubic root nonlinearity is applied to the scalogram $\mathbf{U}_1 x$ before computing amplitude modulations, in order to imitate loudness compression in the cochlea. Second, the final translation-invariant representation is not restricted to average values of $\mathbf{U}_1 x$ and $\mathbf{U}_2 x$, but also entails higher-order statistical moments, i.e. variance, skewness, and kurtosis of $\mathbf{U}_1 x$, as well as local cross-correlations in $\mathbf{U}_2 x$ between coefficients of neighboring γ or γ_2.

Contrary to average values, higher-order moments are unstable to time warps, which would make them unsuitable features for classification. In the sequel, we proceed to show that time-frequency scattering synthesizes textures of comparable or better quality than the McDermott and Simoncelli's representation, with the additional property of Lipschitz-stability to the action of small time warps.

3.3.2 *Gradient descent*

In this subsection, we describe the algorithm used to re-synthesize auditory textures from a target in the space of averaged scattering coefficients.

Initialization

The reconstructed signal is initialized as random Gaussian noise whose power spectral density matches the Fourier transform modulus $|\hat{x}(\omega)|$ of the target $x(t)$. This is achieved by randomizing the phases of $\hat{x}(\omega)$. For $\varphi(\omega)$ a process of independent random variables following a uniform distribution in the interval $[0; 2\pi]$, the reconstructed signal is initialized in the Fourier domain as

$$\hat{y}^{(0)}(\omega) = |\hat{x}(\omega)| \times \exp(\mathrm{i}\varphi(\omega)),$$

and then brought to the time domain by inverse Fourier transform.

Momentum

Denoting by $y^{(n)}(t)$ the reconstruction at step n, the iterative reconstruction procedure is

$$y^{(n+1)}(t) = y^{(n)}(t) + u^{(n)}(t), \tag{3.21}$$

where the additive update $u^{(n)}(t)$ at iteration n is defined recursively as

$$u^{(n+1)}(t) = m \times u^{(n)}(t) + \mu^{(n)} \times \nabla E(y^{(n)})(t). \tag{3.22}$$

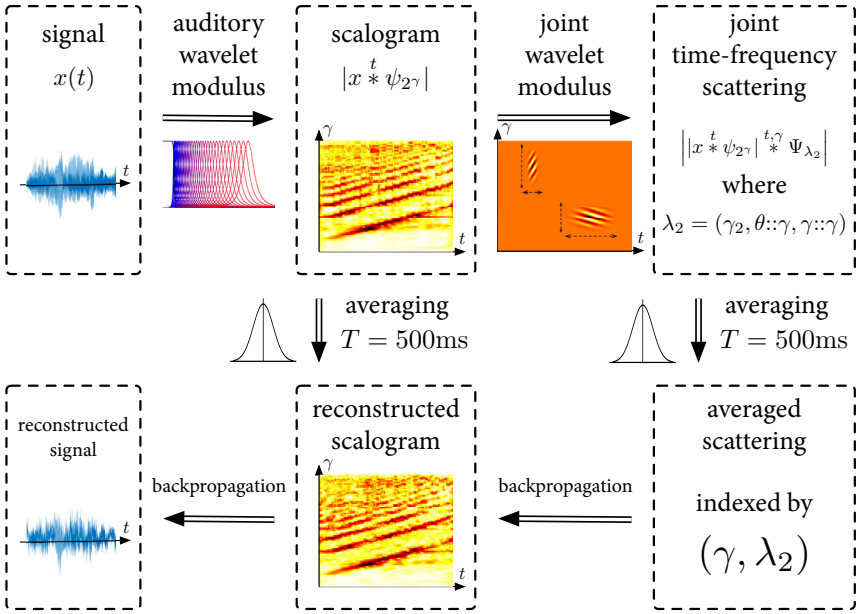


Figure 3.4: Gradient backpropagation of scattering coefficients.

In the equation above, the constant $m > 0$ is the amount of momentum, $\mu^{(n)}$ is the learning rate at iteration n, and $\nabla E(\boldsymbol{y}^{(n)})(t)$ is the gradient of the error functional E with respect to the signal $\boldsymbol{y}^{(n)}$. In all subsequent experiments, the momentum term is fixed at $m = 0.9$. The purpose of momentum is to dampen oscillations of $\boldsymbol{y}^{(n)}$ across iterations within the convex vicinity of the local minimum.

Gradient backpropagation

Like deep neural networks, scattering networks consist of the composition of linear operators (wavelet transforms) and pointwise nonlinearities (complex modulus). Therefore, the gradient $\nabla E(\boldsymbol{y}^{(n)})(t)$ can be obtained by composing the Hermitian adjoints of these operators in the reverse order as in the direct scattering transform. This method, known as gradient backpropagation (Rumelhart, Hinton, and Williams, 1986), is described graphically in Figure 3.4. We refer to Bruna and Mallat (2013a) on the application of gradient backpropagation to scattering networks.

Bold driver heuristic

The sequence of learning rates $\mu^{(n)}$ is initialized at $\mu^{(0)} = 0.1$, and modified at every step to ensure that the error diminishes between iteration. If $E(\boldsymbol{y}^{(n+1)}) < E(\boldsymbol{y}^{(n)})$, $\mu^{(n)}$ is increased by 10% and the update is confirmed; otherwise, $\mu^{(n)}$ is decreased by 50% and the update is retracted. This learning rate policy is known as the "bold driver" heuristic.

3.3.3 *Results*

In this subsection, we set up a qualitative benchmark of audio reconstructions with various architectures, including temporal scattering and time-frequency scattering. We show that the fidelity of time-frequency scattering, unlike temporal scattering, is on par with the state of the art in the domain.

Our benchmark consists of four sounds with different perceptual properties: bubbling water in a pan (3.5a, left); skylark chirp (3.5a, right); spoken English (3.6a, left); and congas (3.6a, right). We compare four invariant representations: averaged scalogram $\mathbf{S}_1 \boldsymbol{x}(t, \gamma)$ only; $\mathbf{S}_1 \boldsymbol{x}(t, \gamma)$ supplemented with temporal scattering $\mathbf{S}_2 \boldsymbol{x}(t, \gamma, \gamma_2)$; $\mathbf{S}_1 \boldsymbol{x}(t, \gamma)$ supplemented with time-frequency scattering $\mathbf{S}_2 \boldsymbol{x}(t, \gamma, \gamma_2, \gamma{::}\gamma, \theta{::}\gamma)$; and the representation of McDermott and Simoncelli (2011).

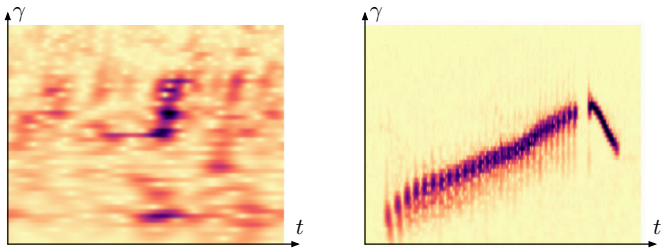
In the case of bubbling water, the reconstruction obtained with only the averaged scalogram lags behind the other three, which compare about equally. This shows that computing temporal modulations is necessary to recover the transientness of non-Gaussian textures, such as those found in environmental sounds.

In the case of the skylark chirp, only time-frequency scattering is able to retrieve the chirp rates of the original sound, and produce coherent patterns in the time-frequency plane. Yet, we must acknowledge that the short silent region between the offset of the ascending chirp and the onset of the descending chirp is absent from the reconstruction. Instead the ascending chirp and the descending chirp cross each over. This is because they are mapped to distinct second-order scattering paths $\lambda_2 = (\gamma_2, \gamma{::}\gamma, \theta{::}\gamma)$.

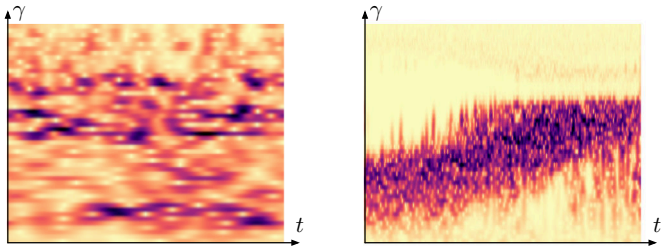
In the case of spoken English, the reconstructed signal from temporal scattering is unintelligible, as it shifts voiced partials back and forth and does not segregate vowels with respect to consonants. On the other hand, the reconstruction obtained from time-frequency remains barely intelligible. This shows the importance of capturing the joint time-frequency structure of patterns in the scalogram, thus segregating chirped harmonic spectra (vowels) from broadband stationary noise (consonants). The reconstruction of McDermott and Simoncelli (2011) cannot be evaluated fairly because it results from the analysis of the full sentence.

Lastly, in the case of congas, the temporal scattering transform fails to recover the synchronicity of impulses across wavelet subbands. However, both time-frequency scattering and the representation of McDermott and Simoncelli (2011) produce well-localized onsets, despite the presence of audible artifacts.

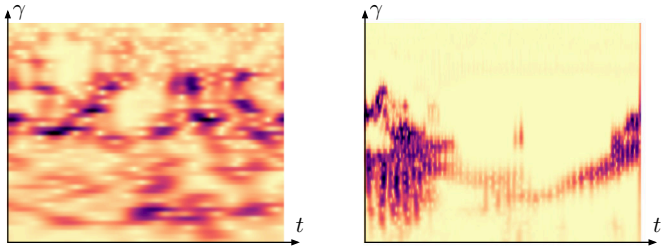
It remains to be established how these methods perform in the more challenging context of re-synthesizing polyphonic music. Section 4.4 will show that spiral scattering provides a slight improvement with respect to time-frequency scattering in this regard.



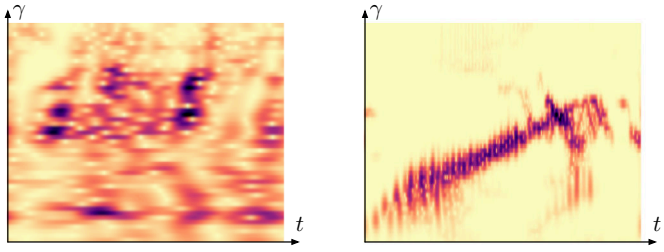
(a) Original.



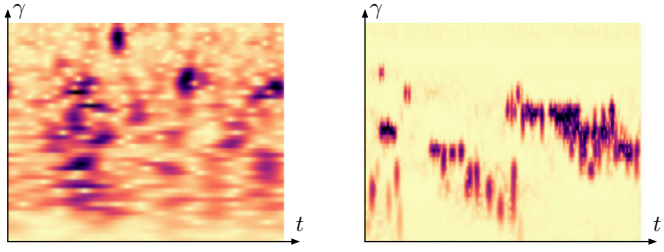
(b) Averaged scalogram.



(c) Averaged scalogram and temporal scattering.

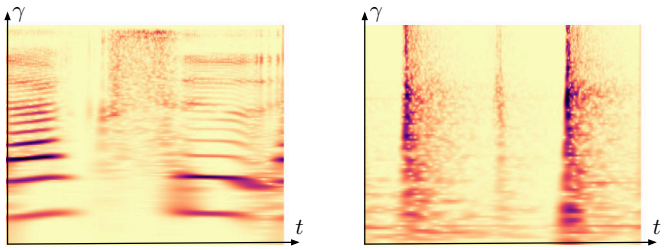


(d) Averaged scalogram and time-frequency scattering.

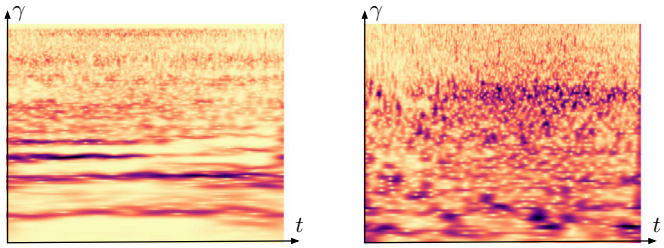


(e) McDermott and Simoncelli's representation.

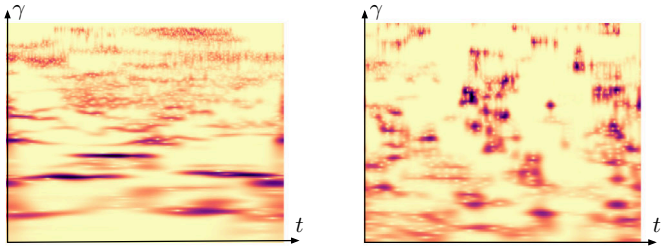
Figure 3.5: Audio texture synthesis.



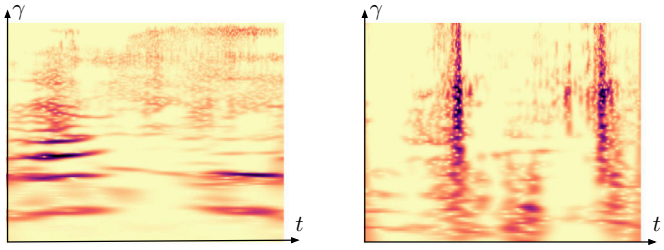
(a) Original.



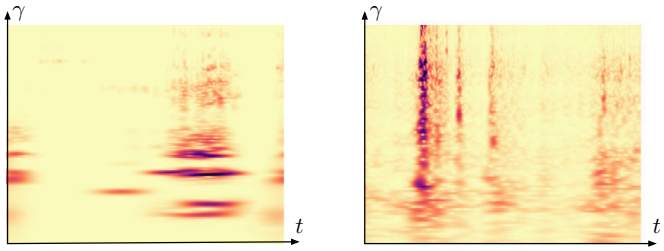
(b) Averaged scalogram.



(c) Averaged scalogram and temporal scattering.



(d) Averaged scalogram and time-frequency scattering.



(e) McDermott and Simoncelli's representation.

Figure 3.6: Audio texture synthesis (bis).

3.3.4 *Creative applications*

The original purpose of this section is to assess the re-synthesis capabilities of time-frequency scattering with respect to existing invariant representations. However, the algorithm above can also be used to create new sounds. In this subsection, we present *FAVN*, a collaborative project undertaken with composer Florian Hecker. As explained in the program notes of the performance, which are partially reproduced hereinafter, *FAVN* takes its roots in Stéphane Mallarmé's poem *L'après-midi d'un faune* (1876) as well as Debussy's *Prélude* (1894).

Protocol

A preliminary version of the piece is composed by means of conventional computer music tools, and mixed down to three monophonic channels. Each channel is a continuous audio stream of duration equal to 17 minutes, segmented into 47 blocks of duration equal to 21 seconds.

Then, each block of each channel is reconstructed independently with the algorithm described previously: once initialized with a Gaussian noise, the reconstruction is updated by gradient descent with a bold driver learning rate policy. The algorithm is stopped after 50 iterations. Every iteration of every block is recorded and sent to the composer.

At the time of the performance, the composer begins by playing the first iteration of the first block, and progressively moves forward in the reproduction of the piece, both in terms of compositional time (blocks) and computational time (iterations).

Figure 3.7 shows the scenography of *FAVN*, as premiered at the Alte Oper in Frankfurt on October 5th, 2016.

Settings

The first-order, auditory filter bank is a nonstationary Gammatone transform with a maximum quality factor $Q_{\max}$ equal to 12 and a maximum time scale $T_{\max}$ of 93 ms. At the second order of the time-frequency scattering transform, both wavelets along time and log-frequency are Gabor wavelets with a quality factor equal to 1.

The averaging size of the low-pass filter $\boldsymbol{\phi}_T(t)$ is set to $T = 188$ ms at both orders of the scattering transform. This duration has been chosen in agreement with the composer by a process of trial and error. We found that setting T to 93 ms or lower led to a convergence rate that was too fast to be useful for a creative application. Conversely, setting T to 372 ms or higher led to the appearance of undesirable artifacts in the reconstruction, even after the local optimum has been reached.

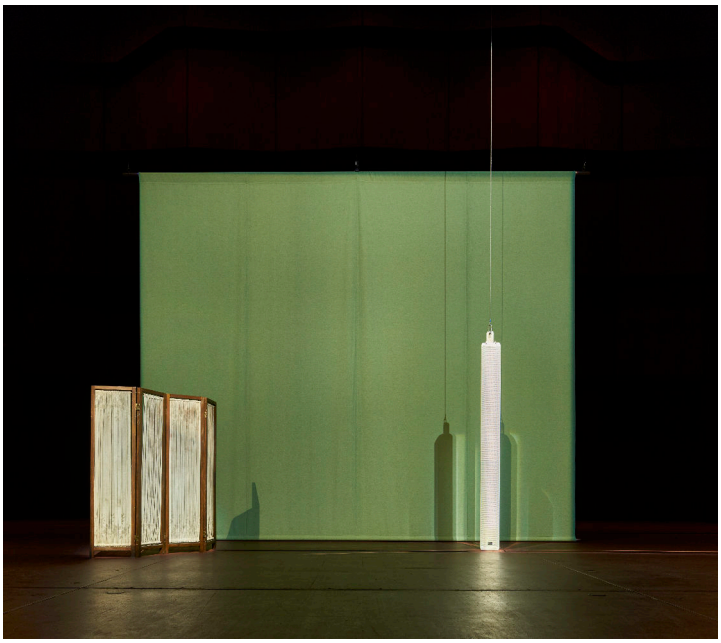


Figure 3.7: The scenography of *FAVN*, by Florian Hecker. This computer music piece was premiered at the Alte Oper in Frankfurt on October 5th, 2016.

Results

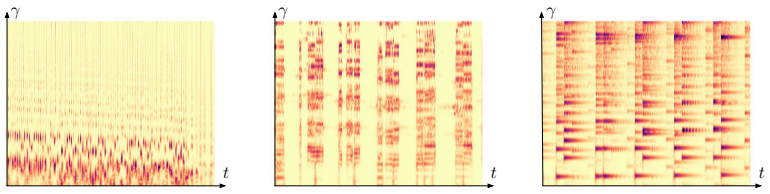
The scalograms of three excerpts of FAVN, all drawn from channel A, are shown in Figure 3.8. The left column shows a texture of harmonic sounds with fast amplitude modulations and a globally descending pitch. The middle column shows a broadband noise with intermittent activity. The right column shows a polyphonic mixture of harmonic sounds with sharp attacks and a slow damping.

The algorithm is initialized with a weak-sense stationary process whose power spectral density matches the target signal, but which does not present any intermittency. As iterations go by, time-frequency patterns emerge in contrast with stationary regions. After about 50 iterations, the algorithm converges to a local optimum which is perceptually close to the original.

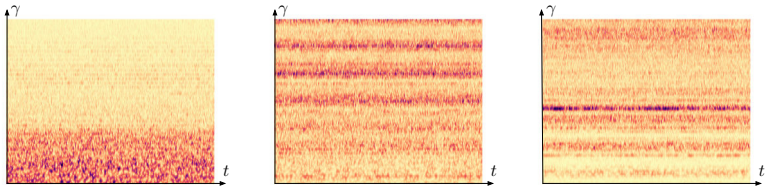
An excerpt from the program notes

We reproduce here an excerpt of the program notes, written by philosopher Robin MacKay for the premiere of *FAVN* in Frankfurt.

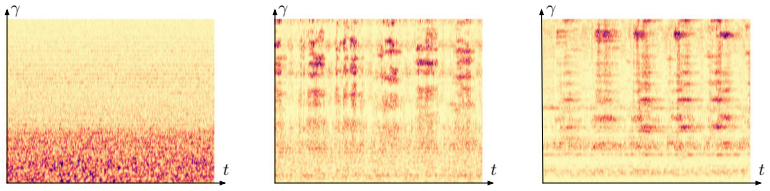
> FAVN also folds Mallarmé's insistence on the impossibility of cataloguing the idea in plain prose onto Hecker's concern with the ways in which sound is analytically coded, in particular focusing on the concept of timbre — certainly a master pertinent to the work of Debussy, renowned as the first composer of *colors* rather than melodies, themes,



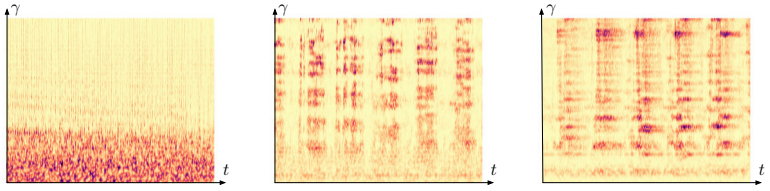
(a) Original.



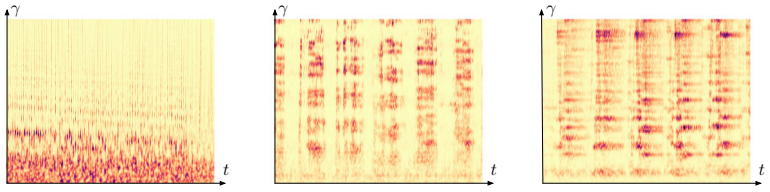
(b) Iteration 1.



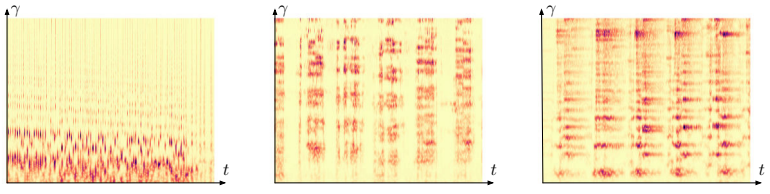
(c) Iteration 11.



(d) Iteration 21.



(e) Iteration 31.



(f) Iteration 50.

Figure 3.8: Three excerpts of *FAVN*, by Florian Hecker. The top row shows the original composition. The following rows shows the re-synthesis of the original composition from its time-frequency scattering coefficients, at various iterations.

> and harmonies, and whose mastery of timbre is magnificently evident in the burnished golds and sunlit verdancies of the Prélude. 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.

Other pieces

Since the creation of *FAVN*, Florian Hecker has used our software to produce four new pieces.

First, *Modulator (Scattering Transform)* is a remix of the electronic piece *Modulator* from 2012, obtained by retaining the 50^{th} iteration of the gradient descent algorithm. It will be released in a stereo-cassette format in early 2017 by Edition Mego.

Secondly, *Modulator (Scattering Transform)* was produced as a 14-channel piece for the "Formulations" exhibition at the Museum für Moderne Kunst in Frankfurt, on display from November 26th, 2016 until February 5th, 2017. In this multichannel piece, 14 loudspeakers in the same room play a different iteration number of the reconstruction algorithm, chosen at random. As a result, the visitor can figuratively walk through the space of iterations while the piece unfolds.

Thirdly, *Experimental Palimpsests* is an 8-channel remix of the electronic piece *Palimpsest* from 2004, obtained by the same procedure. *Experimental Palimpsests* was premiered at the Lausanne Underground Film Festival on October 19th, 2016, in collaboration with artist Yasunao Tone.

Fourthly, *Inspection* is a binaural piece for synthetic voice and electronics, played live at the Maida Vale studios in London and broadcasted on BBC 3 on December, 3rd, 2016.

3.4 APPLICATIONS TO ACOUSTIC SCENE CLASSIFICATION

In this section, we use scattering representations to address two different audio classification problems: environmental sound classification and acoustic scene classification. The former is a source identification problem at a time scale of a few seconds, whereas the latter consists in "classifying environments from the sound they produce" (Barchiesi et al., 2015). Both these problems are rooted in acoustic scene analysis, and are currently addressed by a combination of feature engineering and feature learning (Eghbal-Zadeh et al., 2016).

First of all, we review the related work in the field of environmental sound classification. Secondly, we present the three datasets on which we will evaluate scattering representations: UrbanSound8k, ESC, and DCASE 2013. Thirdly, we discuss the importance of logarithmic compression before feeding scattering coefficients to a support vector machine classifier. Fourthly, we report state-of-the-art results in environmental sound classification and competitive results in acoustic scene classification.

3.4.1 *Related work*

In this subsection, we organize the state of the art in environmental sound classification into four areas: short-term audio descriptors, such as MFCC; unsupervised feature learning, such as matching pursuit or spherical k-means; amplitude modulation features, such as temporal scattering; and deep convolutional networks.

Short-term audio descriptors

The earliest methods for the supervised classification of acoustic scenes rely on short-term audio descriptors, such as mel-frequency cepstral coefficients (MFCC) and their first-order temporal derivatives, spectral shape descriptors, and zero-crossing rate. These descriptors are computed over a typical time scale of $T = 25\,\text{ms}$, and subsequently aggregated through time by means of a Gaussian mixture model (GMM) classifier. We refer to Chachada and Kuo (2014) for a recent survey of the state of the art in this domain.

Unsupervised learning

It has been argued that acoustic scenes are perceptually organized as a "skeleton of events over a bed of texture" (Nelken and Cheveigné, 2013), that is, few time-frequency patterns superimposed over a stationary background. According to this paradigm, it is natural to search for a sparse representation of audio signals which reveals the saliency of acoustic events. Chu, Narayanan, and Kuo (2009) have shown that the activations found by the matching pursuit algorithm (Mallat and

Zhang, 1993) outperform mel-frequency cepstral coefficients (MFCC) in a task of environmental sound classification.

More recently, Stowell and Plumbley (2014) and Salamon and Bello (2015b) have shown that running the spherical k-means clustering algorithm over a large quantity of data allows to build a codebook of mid-level features which can be interpreted as salient events.

Amplitude modulation features

Another perspective on the problem of segregating events from a textured background is brought by the design of amplitude modulation features. Indeed, convolving a scalogram with a modulation filter bank of wavelets tends to decorrelate the contributions of salient events across different time scales.

For example, the representation of McDermott and Simoncelli, originally intended for texture synthesis, was successfully used by Ellis, Zeng, and McDermott (2011) in a task of soundtrack classification.

Baugé et al. (2013) have used the separable time and frequency scattering transform

$$\mathbf{S}_2\boldsymbol{x}(t,\gamma,\gamma_2,\gamma::\gamma) = \left|\left||\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^\gamma}| \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}}\right| \overset{t}{*} \boldsymbol{\phi}_T \overset{\gamma}{*} \boldsymbol{\psi}_{2^{\gamma::\gamma}}\right| (t),$$

introduced by Andén and Mallat (2014), for environmental sound recognition and shown that it outperforms the temporal scattering transform.

Recently, a combination of temporal scattering and spherical k-means has shown to improve the accuracy of urban sound classification with respect to unsupervised learning on scalogram features, while allowing to rely on a smaller codebook (Salamon and Bello, 2015a).

Deep convolutional networks

The latest methods for environmental sound classification rely on supervised representation learning. In particular, the state of the art in the ESC-50 dataset, which will be presented in the next subsection, is held by a deep convolutional network (ConvNet) (Piczak, 2015a,b). Following this success, the 2016 edition of the workshop on detection and classification of acoustic scenes and events (DCASE) has seen a surge of deep convolutional networks (Bae, Choi, and Kim, 2016; Lidy and Schindler, 2016; Valenti et al., 2016).

The main drawback of deep learning techniques is that they need a large quantity of annotated data to converge to a meaningful representation (Le Cun, Bengio, and Hinton, 2015). As of today, the quantity of available data for environmental sound classification is of the order of a few hours, that is, two or three orders of magnitude below automatic speech recognition. Recent work on data augmentation techniques, such as time stretching, pitch shifting, dynamic range

compression, and addition of background noise, has mitigated this issue (Salamon and Bello, 2016).

3.4.2 *Datasets*

In this subsection, we describe the four datasets that will be used to compare algorithms. The former two, UrbanSound8k and ESC-50, consist of short environmental sounds, whereas the latter two, DCASE 2013 and DCASE 2016, consist of acoustic scenes.

UrbanSound8k

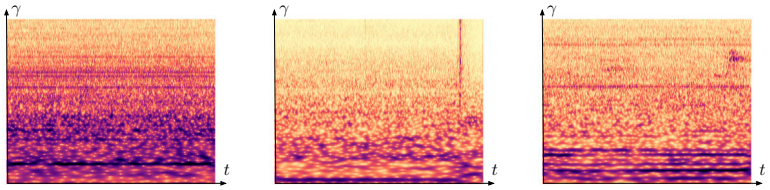
The UrbanSound8k dataset contains 8732 labeled sound excerpts of urban sounds from 10 classes: air conditioner, car horn, children playing, dog bark, drilling, engine at idle, gun shot, jackhammer, siren, and street music. These classes were selected from a large taxonomy of urban sounds because they were ranking among the most frequent causes of noise complaints in the city of New York (Salamon, Jacoby, and Bello, 2014). As such, they are guaranteed to have a strong perceptual salience.

In UrbanSound8k, the number of samples per class varies between 400 and 1000. The duration of these sounds varies from one second to four seconds. To obtain a constant duration of three seconds, we pad shorter sounds with silence and trim longer sounds.

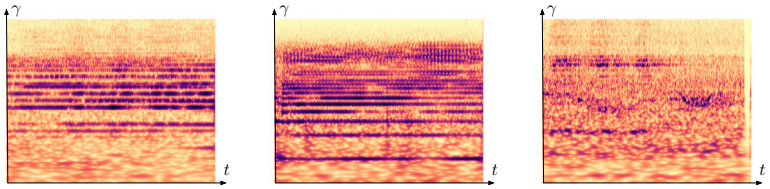
Scalograms of randomly chosen samples for each class are presented in Figures 3.9 and 3.10. We observe that there is a large intra-class variability across samples, which makes the problem difficult. However, we also notice the presence of distinctive time-frequency patterns for each class. For example, jackhammers are singled out by slow impulses, which appear as vertical edges in the scalogram. In contrast, police sirens are modeled by an additive model of exponential chirps, thus resulting in long, diagonal edges. These observations suggest that capturing spectrotemporal modulations in the scalogram is an efficient way to discriminate urban sounds.

ESC-50

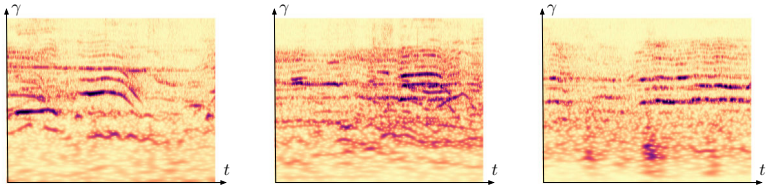
The ESC-50 dataset contains 2000 labeled environmental recordings equally balanced between 50 classes (Piczak, 2015b). Both UrbanSound8k and ESC-50 were assembled by downloading specific categories of sounds from FreeSound (Font, Roma, and Serra, 2013), a freely accessible, collaborative database of audio samples. The human accuracy in ESC-50 varies between 35% (wind, washing machine) and 99% (baby crying, dog, glass breaking).



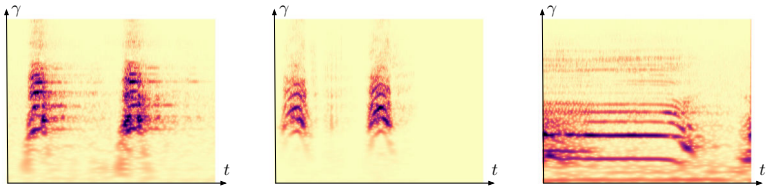
(a) Air conditioner.



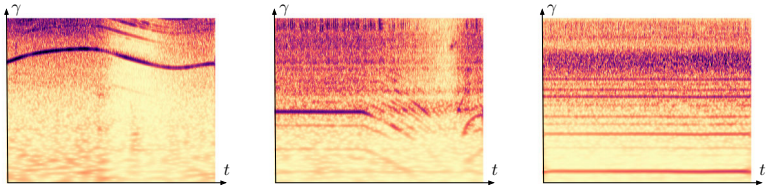
(b) Car horn.



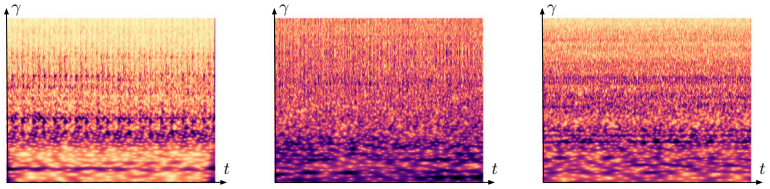
(c) Children playing.



(d) Dog barks.



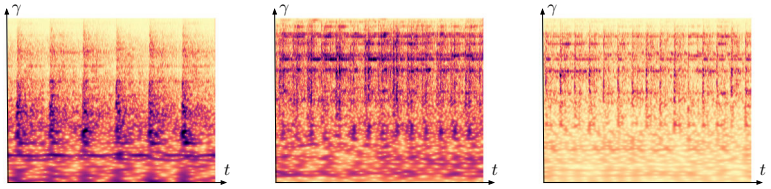
(e) Drilling.



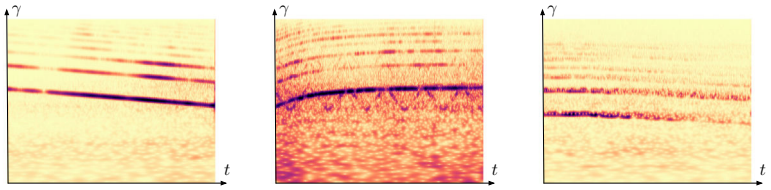
(f) Engine at idle.

Figure 3.9: The UrbanSound8k dataset. Samples from the same row belong to the same class.

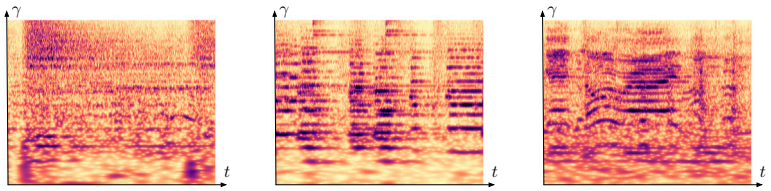
(a) Gunshot.



(b) Jackhammer.



(c) Siren.



(d) Street music.

Figure 3.10: The UrbanSound8k dataset (bis). The UrbanSound8k dataset. Samples from the same row belong to the same class.

DCASE 2013

The DCASE 2013 dataset of acoustic scene classification consists of two parts, namely a public and a private subset, each made of 100 recordings of 30 seconds, evenly divided into 10 classes (Stowell et al., 2015). To build the DCASE 2013 dataset, three different recordists visited a wide variety of locations in Greater London over a period of several months. In order to avoid bias, all recordings were made under moderate weather conditions, at varying times of day and week, and each recordist recorded every scene type. As a consequence, and despite its small size, DCASE 2013 enjoys a challenging intra-class diversity.

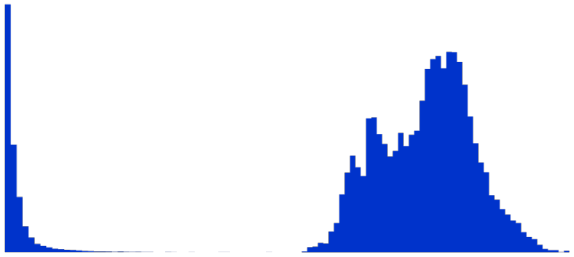
3.4.3 *Methods*

In this subsection, we address the problem of engineering features for the classification of environmental sounds and acoustic scenes. First, we propose a novel data augmentation procedure to encode invariance to the azimuthal position of the recordist in binaural audio. Secondly, we advocate for logarithmic compression of scattering coefficients in order to comply with an assumption of Gaussianity. Thirdly, we compare early integration, which consists in averaging scattering coefficients through time in feature space before training the classifier, versus late integration, which consists in aggregating the short-term outputs of the classifier in decision space.

Binaural data augmentation

Most acoustic scene datasets are recorded according to a binaural protocol, i.e. with a pair of in-ear microphones (Wang and Brown, 2006). This protocol provides a realistic description of the spatial auditory environment, as it reproduces the natural listening conditions of humans. In particular, the interaural level difference (ILD) between left and right channel conveys the approximate azimuth of each sound source that make up the scene with respect to the listener (Blauert, 2004). Yet, the location of the recordist may vary across instances of the same class. Therefore, any global rotations of all azimuths, corresponding to a rotation of the head of the recordist, should be discarded as a spurious source of variability in the observed data.

For classification purposes, averaging the left and right channels into a monophonic signal is by far the most widespread approach, because the resulting monophonic signal is approximately invariant to azimuth. However, it favors the center of the scene while attenuating lateral cues. Moreover, sources that are on distinct azimuths get mixed, which may cause bias at training time. Instead, so as to leverage stereophonic data, we perform multiple combinations of the left



(a) Before logarithmic compression. (b) After logarithmic compression.

Figure 3.11: Statistical effects of logarithmic compression. Histogram of first-order scattering coefficients $\mathbf{S}_1 x(\gamma)$ with $2^\gamma = 300\,\text{Hz}$, before and after logarithmic compression.

channel $x_L(t)$ and the right channel $x_R(t)$ into monophonic signals $x_\alpha(t)$ according to the equation

$$x_\alpha(t) = \frac{1+\alpha}{2} x_L(t) + \frac{1-\alpha}{2} x_R(t)$$

where alpha is a panoramic mixing weight, taking values between -1 and $+1$, which could be deterministic of random. In following experiments, we deterministically set α to the values -1, $-\frac{1}{2}$, 0, $+\frac{1}{2}$, and $+1$. This is a form of data augmentation, because 5 monophonic signals are derived from every binaural recording in the training set. At test time, only the center mix ($\alpha = 0$) is used to predict the class.

Logarithmic compression

Many algorithms in pattern recognition, including nearest neighbor classifiers and support vector machines, tend to work best when all features follow a standard normal distribution across all training instances. Yet, because of the complex modulus nonlinearity, scattering coefficients are nonnegative by design. It appears empirically that their distribution is skewed towards the right, which means that the tail towards greater values is longer than the tail towards lower values. However, skewness can be reduced by applying to all coefficients a pointwise concave transformation, e.g. logarithmic. Figure 3.11 shows the distribution of an arbitrarily chosen scattering coefficient over the DCASE 2013 dataset, before and after logarithmic compression.

The application of the pointwise logarithm to magnitude spectra is ubiquitous in audio signal processing, and is found for instance in mel-frequency cepstral coefficients — see subsection 2.3.3. Indeed, it is corroborated by the Weber-Fechner law in psychoacoustics, which states that the sensation of loudness is roughly proportional to the logarithm of the acoustic pressure in the outer ear. We must also recall that the measured amplitude of sound sources often decays polynomially with the distance to the microphone, which is a spurious factor of variability to the task of urban sound classification. Loga-

rithmic compression can linearize this dependency, which arguably facilitates the construction of a powerful invariant at the classifier stage.

Given a task of musical genre recognition, Andén and Mallat (2014) have advocated for the renormalization of second-order scattering coefficients $\mathbf{S_2}\boldsymbol{x}(t, \gamma_1, \gamma_2)$ by the corresponding first-order scattering coefficients $\mathbf{S_1}\boldsymbol{x}(t, \gamma_1)$, as it provably decorrelates the amplitudes of their activations. Interestingly, taking the logarithm of renormalized coefficients would yield

$$\log \frac{\mathbf{S_2}\boldsymbol{x}(t, \gamma_1, \gamma_2)}{\mathbf{S_1}\boldsymbol{x}(t, \gamma_1)} = \log \mathbf{S_2}\boldsymbol{x}(t, \gamma_1, \gamma_2) - \log \mathbf{S_1}\boldsymbol{x}(t, \gamma_1).$$

i.e. a linear combination of the logarithms of first- and second-order coefficients. Therefore, the theoretical insight brought by Andén and Mallat (2014) in favor of renormalized scattering also applies to log-scattering up to a linear transformation in feature space, to which affine classifiers are not sensitive.

Early and late integration

Owing to the scarcity of salient events in many natural scenes, fine-grained classification is only made possible by integrating over a long temporal context. Indeed, whereas a few seconds are often sufficient to recognize a speaker, a musical instrument, or a genre, it may require up to 30 seconds to disambiguate two classes of acoustic scenes which share part of their semantic content, e.g. a train from a subway station or a quiet street from a park. Depending on whether aggregation is performed in feature space or in decision space, the corresponding method is referred to as early or late integration.

A straightforward application of early integration consists in summarizing the multivariate time series of scattering coefficients over the full duration of the acoustic scene by only storing their average values. Going back to Subsection 2.2.1, this is equivalent to increasing the support T of the low-pass filter $\boldsymbol{\phi}_T(t)$ up to infinity.

Conversely, a late integration scheme relies on probabilistic assignments $\mathbb{P}[y|\mathbf{S}\boldsymbol{x}(t, \lambda)]$ over short-term windows of duration T, which are subsequently aggregated to produce a final decision

$$\hat{y} = \arg\max_y \rho\left(\{\mathbb{P}[y|\mathbf{S}\boldsymbol{x}(t, \lambda)]\}_t\right),$$

where $\hat{y}$ is the estimated class label and ρ is a reduction function, such as sum, product, or majority vote.

The major drawback of early integration is that it drastically reduces the number of training instances at the classifier stage, down to one per acoustic scene. In the context of the DCASE 2013 dataset, this corresponds to merely 8 training instances per class, and 80 instances overall, hence an increase in variance in statistical estimation

and a risk of overfitting. On the contrary, a late integration scheme for $T = 188$ ms would yield 128 instances per acoustic scene, resulting in 10240 instances overall. However, many of these instances may be silent or lack any salient properties of the class they are assigned to, hence an increase in bias and a risk of underfitting. Moreover, when applying late integration, the classifier optimizes accuracy at the frame level, not at the scene level, so its prediction might not yield optimal training error under the final evaluation metric.

In short, early and late integration methods lie at opposite ends of the bias-versus-variance statistical tradeoff. We refer to Joder, Essid, and Richard (2009) for a comprehensive review of this problematic in the context of musical instrument recognition. In this dissertation, early integration refers to an averaging of the features and late integration refers to a majority voting of the predictions.

Support vector machines

A support vector machine (SVM) is a large-margin supervised classifier which finds the optimal separating hyperplane between two classes of samples in feature space (Cortes and Vapnik, 1995). In a multiclass setting, the decision is obtained by aggregating pairwise classifications (Hsu and Lin, 2002). We use a Gaussian kernel and find the optimal parameters σ (variance of the Gaussian) and C (slack variable penalization) by cross-validation. All numerical experiments rely on the LIBSVM package (Chang and Lin, 2011).

3.4.4 *Discussion*

UrbanSound8k dataset

Classification results on the UrbanSound8k dataset are charted in Table 3.1. The baseline consisting of MFCC audio descriptors and a support vector machine (SVM) classifier reaches an average miss rate of 46% across all classes.

The current state of the art, 26.1%, is held by class-conditional spherical k-means applied to time-frequency patches of PCA-whitened log-mel-spectrograms, associated with a random forest classifier (Salamon and Bello, 2015b).

Temporal scattering reaches an average miss rate of 27.1%, a figure on par with the state of the art. Time-frequency scattering reaches an average miss rate of 20.6%, thus outperforming the state-of-the-art by two standard deviations. Performances without logarithmic compression are considerably worse, and are thus not reported.

ESC-50 dataset

Classification results on the ESC-50 dataset are charted in Table 3.2. The baseline reaches an average miss rate of 56% across all classes.

Table 3.1: Classification results on the UrbanSound8k dataset.

features	classifier	average miss rate
MFCC	SVM	46.0
scalogram + spherical k-means	random forest	26.1
temporal scattering	SVM	27.1 ± 3.0
time and frequency scattering	SVM	23.5 ± 4.0
time-frequency scattering	SVM	20.6 ± 3.5

The current state of the art, 35.5%, is held by deep convolutional networks, which are trained on mel-frequency spectrograms with logarithmic compression, as well as their first-order temporal derivatives (Piczak, 2015a). In the state-of-the-art implementation, the size of the receptive fields is chosen to be slightly less than the number of filters, thus providing a small amount of covariance with frequency transposition.

With an average miss rate of 40.5%, temporal scattering falls behind the state of art. Yet, separable time and frequency scattering (30.8%) and time-frequency scattering (28.0%) both outperform deep convolutional networks. This experiment shows that the use of a multiresolution scheme in the design of spectrotemporal descriptors can outperform deep representations which are learned on single-resolution inputs, such as mel-frequency spectrograms.

Table 3.2: Classification results on the ESC-50 dataset.

features	classifier	average miss rate
MFCC	SVM	56.0
scalogram	ConvNet	35.5
temporal scattering	SVM	40.5 ± 4.4
separable time and frequency scattering	SVM	30.8 ± 4.4
joint time-frequency scattering	SVM	28.0 ± 3.8

DCASE 2013 dataset

Classification results on the DCASE 2013 dataset are charted in Table 3.3. The baseline reaches an average miss rate of 45% across all classes.

Spectrotemporal features based on a single-resolution discretization of the scalogram, such as spectrotemporal receptive fields (STRF, Patil and Elhilali) and histogram of gradients (HoG, Rakotomamonjy and Gasso), respectively reach average miss rates of 42% and 31%, thus performing better than the baseline.

The winners of the DCASE 2013 challenge are Roma et al. (2013), which complemented MFCC with recurrence quantification analysis (RQA) features, and achieved an average miss rate of 24% with an SVM classifier. We refer to Stowell et al. (2015) for an overview of all submitted systems.

In the case of temporal scattering, we find that late integration (28%) outperforms early integration (41%). On the contrary, in the case of time-frequency scattering, we find that early integration (20%) outperforms late integration (23%). In the realm of acoustic scene analysis, this finding suggests that time-frequency scattering is better at segregating auditory information across multiple paths λ_2 than temporal scattering, thus enabling a broader temporal integration. Furthermore, our binaural data augmentation method has a crucial role for artificially increasing the number of examples per class in the case of early integration: it brings the average miss rate from 42% to 41% for temporal scattering and from 32% to 20% for time-frequency scattering.

Since the end of the DCASE 2013 challenge, the state of the art has been improved by Agcaer et al. (2015), who reported an average miss rate of 13%. To achieve this figure, the authors employed label tree embedding (LTE) a supervised method to learn class-label hierarchies. This result shows that acoustic scene classification is not only a matter of engineering discriminative features, but can also be addressed by advanced machine learning techniques in the semantic space of classes.

DCASE 2016 dataset

With Joakim Andén, we took part in the 2016 edition of the DCASE challenge (Mesaros, Heittola, and Virtanen, 2016). The corresponding dataset consists of 1170 recordings of 30 seconds, evenly divided into 15 classes. As such, it is one order of magnitude larger than the DCASE 2013 dataset, which allows to resort on data-intensive machine learning techniques.

We submitted a system based on binaural data augmentation, temporal scattering with Gammatone wavelets ($T = 743$ ms), logarithmic compression, and support vector machine classification with a linear kernel. In the public dataset, we achieved an accuracy of 79.4%, whereas the baseline was at 70.8%. In the private dataset, we achieved an accuracy of 80.8%, whereas the baseline was at 77.2%. Due to lack of time, time-frequency scattering was not submitted to the challenge.

Table 3.3: Classification results on the DCASE 2013 dataset. See text for details.

features	classifier	integration	miss rate
MFCC	GMM	late	45
STRF	SVM	late	42
HoG	SVM	late	31
MFCC + RQA	SVM	late	24
scalogram	LTE+SVM	late	13
temporal scattering	SVM	early	41
temporal scattering	SVM	late	28
time-frequency scattering	SVM	early	20
time-frequency scattering	SVM	late	23

The best results were reached by Eghbal-Zadeh et al. (2016) (89.7%), who combined engineered features (binaural i-vector) and learned features (deep convolutional networks); and by Bisot et al. (87.7%), who performed supervised nonnegative matrix factorization (NMF) over the scalogram.

In conclusion, scattering transforms may be sufficient to integrate the auditory information in a short environmental sound, in which T is less than one second, but do not match feature learning techniques when applied to a problem of acoustic scene classification, in which T is equal to thirty seconds.

4

SPIRAL SCATTERING

In the previous chapter, we have presented two convolutional operators in the time-frequency domain: temporal scattering and time-frequency scattering. We have shown that time-frequency scattering is more discriminative than temporal scattering, resulting in more accurate texture synthesis and an improved accuracy in audio classification. This is because time-frequency scattering provides a sparse representation of spectrotemporal modulations, which are commonly found in speech, bioacoustics, and acoustic scene analysis.

Yet, musical signals do not only consist of spectrotemporal modulations, but also more intricate patterns, and notably harmonic intervals between partials. The problem of jointly characterizing the locations and amplitudes of these partials, as well as their synchronicity and relative evolution, is at the heart of many tasks in audio classification, including musical instrument recognition.

This chapter introduces a new scattering representation, called spiral scattering, specifically designed for harmonic or quasi-harmonic sounds. Spiral scattering is a composition of three wavelets transforms over the scalogram, respectively applied along time, along log-frequency, and across octaves. The name "spiral" is chosen in reference to the seminal experiments of Shepard (1964) and Risset (1969) in music psychology.

The rest of this chapter is organized as follows. In Section 4.1, we gather arguments in favor of rolling up the log-frequency axis into a spiral which makes one turn at every octave. In Section 4.2, we define the spiral scattering transform. In Section 4.3, we study the spiral scattering coefficients of a nonstationary generalization of the source-filter model, in which both pitch and spectral envelope are deformed through time. In Section 4.4, we re-synthesize music from spiral scattering and show that harmonic intervals are better recovered than with time-frequency scattering. In Section 4.5, we show that spiral scattering outperforms time-frequency scattering and deep convolutional networks in a task of musical instrument recognition.

Part of the content of this chapter has been previously published by Lostanlen and Mallat (2015) and Lostanlen and Cella (2016).

4.1 PITCH CHROMA AND PITCH HEIGHT

The periodicity of musical scales hints for a circular geometry of pitch. In order to reconcile it with the rectilinear geometry of acoustic frequency, the log-frequency axis can be rolled up into a spiral which

makes a full turn at every octave. In polar coordinates, the spiral is parametrized by its angular and radial variables, respectively called pitch chroma and pitch height.

In this section, we address the problem of disentangling pitch chroma from pitch height in the time-frequency domain. Subsection 4.1.1 argues in favor of adopting the octave interval as the circumference of the pitch chroma circle. In particular, we show that the geometry of the Shepard pitch helix can be retrieved from acoustical data by applying a nonlinear dimensionality reduction algorithm. Subsection 4.1.2 is devoted to the construction of Shepard tones, which have a pitch chroma but no definite pitch height. Subsection 4.1.3 presents two spatial models of pitch that disentangle pitch chroma from pitch height, namely the Shepard pitch helix and the Shepard pitch spiral.

4.1.1 *Octave equivalence*

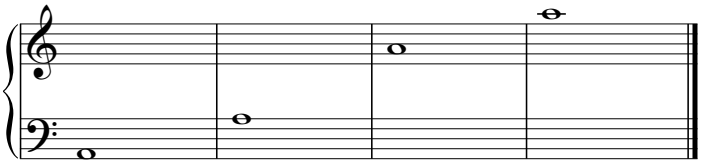
The octave interval corresponds to a multiplication of the fundamental frequency by a factor of two. It is particularly consonant, because raising a harmonic sound by exactly one octave preserves the location of even-numbered partials, whereas is cancels odd-numbered partials. As a result, musical scales consist of a finite number of pitch classes, which are repeated periodically at every octave — a phenomenon known as octave equivalence.

In this subsection, we gather four arguments supporting the adoption of octave equivalence in music signal processing, respectively coming from ethnomusicology, auditory neuroscience, unsupervised learning, and arithmetics. The former two are drawn from the scientific literature, whereas the latter two are novel work.

Argument from ethnomusicology

There is a millennial tradition of music solmization, that is, assigning a distinct symbol to each note in a musical scale. Western music theory has progressively evolved throughout the Middle Ages, and now consists of a chromatic scale of twelve pitches, comprising a diatonic scale of seven pitches denoted by letters from A to G. Whatever be the chosen spacing between semitones, called temperament, the full chromatic scale spans exactly one octave. This range is subsequently expanded by octave transposition, which is figured by appending an integer subscript to the letter, e.g. A_4 for the tuning standard at 440 Hz. Indeed, the harmonic consonance of octave intervals is so strong that it is convenient to assign the same literal to tones that are one octave apart, thus giving a particular status to octaves with respect to other intervals.

It should be remarked that, although the physical notion of acoustic frequency varies rectilinearly from low to high, Western musicians associate two distinct quantities to every musical tone. The former,



A_1 (110 Hz) A_3 (220 Hz) A_4 (440 Hz) A_5 (880 Hz)

Figure 4.1: Octave equivalence. Four musical notes of same pitch chroma (A) and pitch heights ranging from 1 to 4. Observe that the octave interval corresponds to a doubling of the fundamental frequency.

called *pitch class* in music theory or *pitch chroma* in music technology, relates to the location in the chromatic scale, and is commonly described by a letter. The latter, called *pitch height*, relates to the coarse register of the tone, and is described by an integer number — see Figure 4.1. Like acoustic frequency, pitch height is amenable to a rectilinear axis, at it is only bounded by our hearing range. Nevertheless, pitch chroma is commonly represented on a circle, because it wraps around itself with a period of one octave. This fact, known as octave equivalence, is not limited to the Western tradition, but it is also found in many non-Western cultures, including Indian (Pesch, 2009), Chinese (Van Aalst, 2012), and Japanese (Malm, 2000).

Argument from auditory neuroscience

There is an ongoing debate as to whether octave equivalence is innate or acquired. We refer to Burns (1999) for a review. Ethnomusicologists have found some examples of music cultures which ignore the octave interval, thus disputing octave equivalence as a cultural universal (Nettl, 1956). Yet, there is some neurophysiological evidence, based on functional magnetic resonance imaging (fMRI), suggesting that the pitch chroma and pitch height are mapped to distinct regions of the primary auditory cortex of mammals (Briley, Breakey, and Krumbholz, 2013; Warren et al., 2003a,b). In order to demonstrate this, the authors have designed a sawtooth-shaped signal

$$\boldsymbol{x}(t) = \sum_p \frac{1}{p} \cos(2\pi \times p\xi t)$$

of fundamental frequency ξ, with harmonic partials of frequencies $p\xi$ decaying in inverse proportion to the integer p. From the stationary signal $\boldsymbol{x}(t)$, they built two nonstationary signals: first, by an ascending frequency transposition (*glissando*), thus varying pitch chroma and pitch height coherently; secondly, by progressively canceling odd-numbered partials, thus varying pitch height without affecting pitch chroma.

Because the spectral envelope $\omega \mapsto 1/\omega$ is invariant to dilation, frequency transposition of $\boldsymbol{x}(t)$ can be implemented by an exponential

time warp, in the absence of any sound transient — see Section 2.3. For the pitch chroma to make a full circle after some duration T, the time warp is defined as

$$\boldsymbol{\tau}(t) = \frac{T}{\log 2} \times 2^{\frac{t}{T}}.$$

The temporal constant T, typically of the order of one second, is large is front of the pseudo-period $\frac{1}{\xi}$. After first-order Taylor expansion, it appears that the instantaneous frequency of the warped sound

$$\begin{aligned} \mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{x}(t) &= \dot{\boldsymbol{\tau}}(t) \times (\boldsymbol{x} \circ \boldsymbol{\tau})(t) \\ &= 2^{\frac{t}{T}} \times \sum_p \frac{1}{p} \cos\left(2\pi \times \frac{p\xi T}{\log 2} 2^{\frac{t}{T}}\right) \end{aligned} \tag{4.1}$$

is equal to $2^{\frac{t}{T}}\xi$. In particular, the fundamental frequency of $\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{x}(t)$ is equal to 2ξ at $t = T$. Therefore, the chirp rate of the nonstationary signal $\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{x}(t)$ is equal to $\frac{1}{T}$ octaves per second.

Alternatively, the pitch height of $\boldsymbol{x}(t)$ can be progressively raised without affecting the pitch chroma by making the odd-numbered partials $(2p+1)\xi$ progressively vanish throughout the time interval $[0; T]$. For $t > T$, a change of variable $p' = 2p$ in the pseudo-infinite sum of partials yields

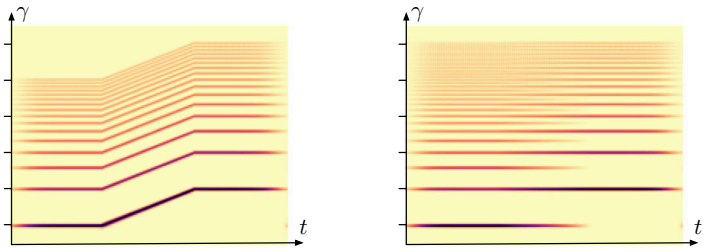
$$\sum_p \frac{1}{2p} \cos(2\pi \times 2p\xi t) = \frac{1}{2} \sum_{p'} \frac{1}{p'} \cos(2\pi \times p' \times (2\xi) \times t),$$

that is, a sawtooth-shaped signal of fundamental frequency 2ξ. Again, the nonstationary signal obtained from $\boldsymbol{x}(t)$ by cancellation of odd-numbered partials have increased in pitch by exactly one octave after the duration T. The scalograms of both signals are shown in Figure 4.2.

Neglecting the topmost octave of their spectra, which lies beyond the audible frequency range, the two signals have identical stationary regions. Furthermore, in the transient regime, the slowness of their amplitude and frequency modulations creates a sensation of temporal continuity. This example illustrates that there are two continuous paths from C_4 to C_5, one circular and one rectilinear. To reconcile them, it is necessary to map acoustic frequencies onto a one-dimensional differentiable manifold in a higher-dimensional vector space. As will be seen in the next section, three-dimensional helixes and two-dimensional spirals are appropriate candidates for such a mapping.

Argument from unsupervised learning

It remains to be established whether octave equivalence can be retrieved from the data without any prior knowledge on ethnomusicol-



(a) Frequency transposition.

(b) Cancellation of odd-numbered partials.

Figure 4.2: Two continuous paths from C_4 to C_5, (a) by frequency transposition or (b) by canceling odd-numbered partials. These sounds have been used as stimuli to demonstrate that pitch chroma and pitch height are mapped to distinct regions of the primary auditory cortex (Warren et al., 2003b). Ticks on the vertical axis denote octave intervals.

ogy or auditory neuroscience. To support this claim, we compute the averaged wavelet scalogram features

$$\mathbf{S}_1 \boldsymbol{x}_n(\gamma) = \int_{-\infty}^{+\infty} |\boldsymbol{x}_n * \boldsymbol{\psi}_{2^\gamma, \frac{Q}{2^\gamma}}|(t)\ \mathrm{d}t$$

of a collection of natural sounds $\{\boldsymbol{x}_n(t)\}_n$ and apply a nonlinear dimensionality reduction algorithm on the cross-correlations between scalogram bins γ, thus visualizing how the trajectory log-frequency axis wraps around itself in dimension three. This experimental protocol is inspired from Le Roux et al. (2007), who demonstrated that the topology of pixels in natural images, i.e. a grid of equidistant points forming square cells, can be retrieved from observing cross-correlations between pixel activations in a dataset of handwritten digits.

We begin by computing the sample means

$$\mu(\gamma) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{S}_1 \boldsymbol{x}_n(\gamma)$$

and standard deviations

$$\sigma(\gamma) = \frac{1}{N-1} \sum_{n=1}^{N} \sqrt{\mathbf{S}_1 \boldsymbol{x}_n(\gamma) - \mu(\gamma)}$$

of every feature across the dataset. Then, we measure the standardized empirical correlations

$$C(\gamma, \gamma') = \frac{\sum_{n=1}^{N} (\mathbf{S}_1 \boldsymbol{x}_n(\gamma) - \mu(\gamma)) \times (\mathbf{S}_1 \boldsymbol{x}_n(\gamma') - \mu(\gamma'))}{\sigma(\gamma) \times \sigma(\gamma')}$$

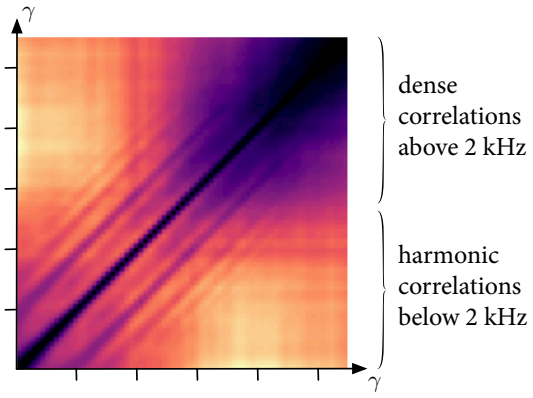


Figure 4.3: Pseudo-distances between averaged scalogram coefficients in the RWC dataset. Darker colors denote smaller distances. Axis ticks denote octave intervals.

between pairs of scalogram bins (γ, γ'). These empirical correlations range between -1 and 1. If we assume their absolute values to be equal to a Gaussian kernel

$$|C(\gamma, \gamma')| = K(\boldsymbol{y}(\gamma), \boldsymbol{y}(\gamma')) = \exp\left(-\|\boldsymbol{y}(\gamma) - \boldsymbol{y}(\gamma')\|_2^2\right),$$

between pairs of points $\boldsymbol{y}(\gamma)$ and $\boldsymbol{y}(\gamma')$, then by defining a self-distance matrix $D(\gamma, \gamma') = \|\boldsymbol{y}(\gamma) - \boldsymbol{y}(\gamma')\|_2$, we obtain the formula

$$D(\gamma, \gamma') = \sqrt{-\log|C(\gamma, \gamma')|}.$$

Figure 4.3 shows the self-distance matrix $D(\gamma, \gamma')$ in the case of the dataset $\{\boldsymbol{x}_n(t)\}_n$ of 6140 isolated notes from 14 instruments, derived from the RWC dataset (Goto et al., 2003). Because adjacent wavelets in the filterbank overlap in frequency, we observe that the main diagonal line is thick. Moreover, in the bottom-left quadrant of the matrix, corresponding to frequencies 2^γ below 2 kHz, we observe sub-diagonals of strong correlations, near cells (γ, γ') of the form

$$\gamma' = \gamma \pm \log_2 p$$

for integer p. In particular, $p = 2$ yields an octave, and $p = 3$ yields an octave plus a perfect fifth. In the top-right quadrant, corresponding to frequencies 2^γ above 2 kHz, correlations are dense. This is because the number of peaks in a harmonic spectrum grows exponentially with log-frequency.

Once the self-distance matrix $D(\gamma, \gamma')$ has been obtained, our aim is to generate a set of points $\{\boldsymbol{v}(\gamma)\}_\gamma$ in a space of dimension $d = 3$ such that distances are preserved, that is,

$$D(\gamma, \gamma') \approx \|\boldsymbol{v}(\gamma) - \boldsymbol{v}(\gamma)\|_2.$$

To do so, we apply a classical algorithm for nonlinear dimensionality reduction named Isomap (Tenenbaum, De Silva, and Langford, 2000).

The core idea behind Isomap is to place the points $v(\gamma)$ on a manifold of dimension d such that the geodesic distance between any pair $(v(\gamma), v(\gamma'))$ is approximated by the shortest path linking them in the neighborhood graph induced by the matrix $D(\gamma, \gamma')$ of pairwise distances. The construction of this neighborhood graph is conditional upon a chosen number of neighbors k for every vertex.

For $k = 1$ or $k = 2$ neighbors, the points $v(\gamma)$ are arranged on a straight line. This is because the two most similar coefficients to a given wavelet bin $\mathbf{S}_1 x(\gamma)$ are its lower and upper neighbors in log-frequency γ, that is, $\mathbf{S}_1 x(\gamma - \frac{1}{Q})$ and $\mathbf{S}_1 x(\gamma + \frac{1}{Q})$. Therefore, as one would expect, the foremost factor of connectedness among wavelet coefficients is consistent with the natural continuity of the frequency variable. However, for $k = 3$ or larger, a new dimension of regularity appears. Due to correlations across octaves, wavelet coefficients $\mathbf{S}_1 x(\gamma)$ are often linked to the coefficient one octave higher or lower, that is, $\mathbf{S}_1 x(\gamma \pm 1)$.

Once embedded in dimension three, the log-frequency axis appears to wrap around itself and forms a helix that makes a full turn at every octave, as shown in Figure 4.4. In cylindrical coordinates, pitch chroma corresponds to an angular variable whereas pitch height corresponds to a rectilinear variable. Subsection 4.1.3 will provide further insight on the helical model of pitch (Deutsch, Dooley, and Henthorn, 2008; Shepard, 1964).

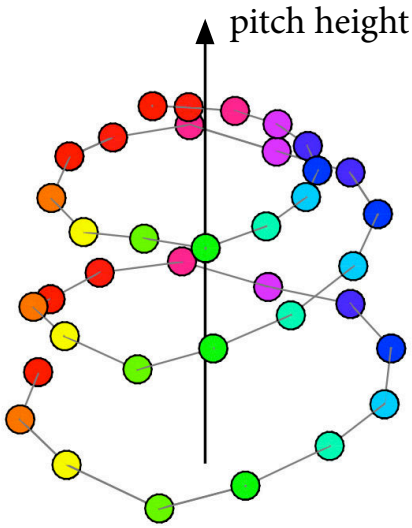


Figure 4.4: Isomap embedding of averaged scalogram features reveals the Shepard pitch helix. The hue of colorful dots denotes their pitch chroma. The solid black line joining adjacent semitones and the vertical axis denoting pitch height were added after running the Isomap algorithm.

This result, however, is limited to datasets of musical instruments and to frequencies below 2 kHz. Above this limit, the log-frequency

axis becomes rectilinear, because octave relationships are no longer relevant. Furthermore, replacing the RWC dataset by the UrbanSound8k dataset produces a rectilinear embedding instead of a helical embedding, because urban sound spectra have less harmonic structure than musical spectra.

Argument from arithmetics

Aside from time-frequency analysis, an arithmetical argument can be devised in favor of octave equivalence in constant-Q filter banks. We consider a discretized filter bank of wavelets, whose center frequencies r^n are in geometric progression with $r > 1$; and a harmonic comb of $(P+1)$ partials, whose dimensionless fundamental frequency is set to 1. Let

$$\mathcal{E}_P(r) = \{n \in \mathbb{N} \mid r^n \in [\![0; P]\!]\}$$

be the set of frequencies intersecting the wavelet filter bank and the harmonic comb. Our aim is to prove that the cardinal of the intersection $\mathcal{E}_P(r)$ is maximized if the scale factor r is the Q^{th} root of 2 for some integer Q.

Let $(n_i)_i$ be the ordered sequence of indices satisfying $r^{n_i} \in \mathbb{N}$. Because $(i \times n_1)$ is a subsequence of $(n_i)_i$, we ascertain that the sequence $(n_i)_i$ is infinite. Leveraging the unicity of the prime number decomposition, it is verified that every integer r^{n_i} is a divisor of $r^{n_{i'}}$ for $i < i'$. By strong recurrence, we deduce that the integer n_i is actually $i \times n_1$. The definition of $\mathcal{E}_P(r)$ rewrites as

$$\mathcal{E}_P(r) = \left\{i \in \mathbb{N} \mid r^{i \times n_1} \leq P\right\}.$$

We then set $Q = n_1$, i.e. the smallest non-null integer such that r^Q be integer. The inequality $r^{i \times Q} \leq P$ is equivalent to $i \leq \log_{r^Q} P$. Therefore, the cardinal of the set $\mathcal{E}_P(r)$ is equal to

$$\operatorname{card} \mathcal{E}_P(r) = 1 + \lfloor \log_{r^Q} P \rfloor .$$

It stems from the above that every r maximizing $\operatorname{card} \mathcal{E}_P(r)$ also maximizes $\lfloor \log_{r^Q} P \rfloor$, and thus minimizes $\log r^Q$ for fixed P. Because r^Q is defined as an integer strictly above 1, $\operatorname{card} \mathcal{E}_P(r)$ is maximal for $r^Q = 2$, i.e. for $r = \sqrt[Q]{2}$. This result is independent of the number of partial waves $(P+1)$ in the harmonic comb.

An interpretation of this result is that adopting an integer number Q of wavelets per octave is the optimal choice of discretization, because it allows to match the octave relationships in harmonic sounds. Writing the center frequencies r^n as $\omega = 2^\gamma$ leads to

$$\log_2 \omega = \log_2 r^n = n \times \log_2 \sqrt[Q]{2} = \frac{n}{Q},$$

i.e a uniform quantization of the log-frequency axis with a step of $\frac{1}{Q}$. In the case $Q = 1$, that is $r = 2$, the log-frequency index $\gamma = \log_2 \omega$ is expressed as an integer number of octaves, and all wavelets coincide with a power-of-two harmonic partial. For $Q > 1$, only integer values γ coincide exactly with a sound partial, whereas non-integer values of γ do not.

In the special case $Q = 12$, that is $r = \sqrt[12]{2}$, the irrational numbers r^{19} and r^{28} happen to be approximately equal to the integers 3 and 5, i.e. the Pythagorician intervals of perfect fifth and major third (Jedrzejewski, 2002). Thus, for applications in musical instrument classification, we choose $Q = 12$ or $Q = 24$ so as to match twelve-tone equal temperament, while striking a good compromise in time-frequency localization.

4.1.2 *Shepard tones*

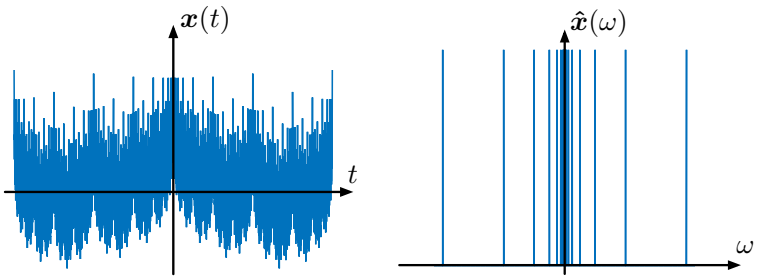
In this subsection, we design a sound, called Shepard tone, by stacking partial waves in octave relationship. We show that Shepard tones are self-similar in the time-frequency domain. Because a Shepard tone has a pitch chroma but no definite pitch height, it can be used to create paradoxical experiments in the cognition of musical pitch.

Self-similarity in the Fourier domain

In order to disentangle pitch chroma from pitch height in perceptual experiments, psychologist Roger Shepard designed a synthetic sound, named Shepard tone, which has a pitch chroma but no definite pitch height (Shepard, 1964). A Shepard tone consists of sine waves in octave relationship, thus covering the full hearing range. Since the octave interval corresponds to a multiplication of the fundamental frequency by a factor of two, the partial waves in a Shepard tone have frequencies of the form $2^\rho \xi$ for $\rho \in \mathbb{Z}$. In practice, the human audible spectrum ranges between 20 Hz and 20 kHz, so it is sufficient to restrict the integer ρ to a finite set of consecutive values. The additive sinusoidal model of the Shepard tone is

$$x(t) = \sum_\rho 2^\rho \cos(2\pi 2^\rho \xi t)$$

Neglecting the boundaries of the audible spectrum, the Shepard tone $x(t)$ is invariant to dilation by a factor of two, as it satisfies the equation $2x(2t) = x(t)$. Therefore, it is a particular case of a self-similar signal, also called fractal signal. Figure 4.5 displays the self-similar structure of the Shepard tone waveform, which is made even clearer in the Fourier domain.



(a) Time domain. (b) Fourier domain.

Figure 4.5: A Shepard tone: (a) time domain; (b) Fourier domain. Observe the self-similar structure of the signal.

Self-similarity in the time-frequency domain

There is a profound connection between fractals and wavelet theory (Bacry, Muzy, and Arnéodo, 1993; Mallat and Hwang, 1992). Indeed, because the wavelets $\boldsymbol{\psi}_{2^\gamma, \frac{Q}{2^\gamma}}(t)$ rely on the dilations of a single wavelet $\boldsymbol{\psi}(t)$, they satisfy a self-similarity equation

$$\boldsymbol{\psi}_{2^\gamma, \frac{Q}{2^\gamma}}(t) = 2\boldsymbol{\psi}_{2^{\gamma+1}, \frac{Q}{2^{\gamma+1}}}(2t)$$

at every frequency 2^γ. By a homothetic change of variable, the self-similarity property of the fractal signal $\boldsymbol{x}(t)$ is readily transferred to the continuous wavelet transform at $t = 0$:

$$\mathrm{CWT}(\boldsymbol{x})(t = 0, \gamma) \quad = \quad \mathrm{CWT}(\boldsymbol{x})(t = 0, \gamma - 1),$$

of which we deduce that the wavelet scalogram of the Shepard tone is periodic over the log-frequency axis at $t = 0$, with a period equal to one octave (Schroeder, 1986). Furthermore, since the quality factor Q of the auditory wavelet $\boldsymbol{\psi}(t)$ is above 1, the filter bank has a sufficient frequential resolution to discriminate partial frequencies $2^\rho\xi$ in the Shepard tone. For lack of any interference between partials, the scalogram $\mathbf{U}_1\boldsymbol{x}(t, \gamma)$ remains constant through time. Therefore, the periodicity equation

$$\mathbf{U}\boldsymbol{x}(t, \gamma) = \mathbf{U}\boldsymbol{x}(t, \gamma \pm \rho) \tag{4.2}$$

is extended to all time instants t and all octaves $\rho \in \mathbb{Z}$.

Chromatic scales of Shepard tones and tritone paradox

From a musical perspective, because any frequency $2^\rho\xi$ could qualify as a fundamental frequency for the Shepard tone, its pitch height is neither high nor low. Yet, Shepard tones can be placed on a chromatic scale, as their pitch chroma is controlled by ξ. Listening experiments demonstrated that, for any given pair of Shepard tones $\boldsymbol{x}(t)$ and $\boldsymbol{x}'(t)$ of respective frequencies ξ and $\xi' > \xi$, the pitch of $\boldsymbol{x}(t)$ is judged

lower than $x'(t)$ if the ratio between ξ' and ξ is between 1 and $\sqrt{2}$, and judged higher if this ratio is between $\sqrt{2}$ and 2 (Shepard, 1964). Therefore, a repeated chromatic scale of Shepard tones seems to ascend or descend endlessly in pitch. This famous auditory illusion has led to the theory of pitch circularity, which emphasizes the fact that the pitch chroma wraps around periodically. The scalogram of a chromatic scale of Shepard tones is shown in Figure 4.6.

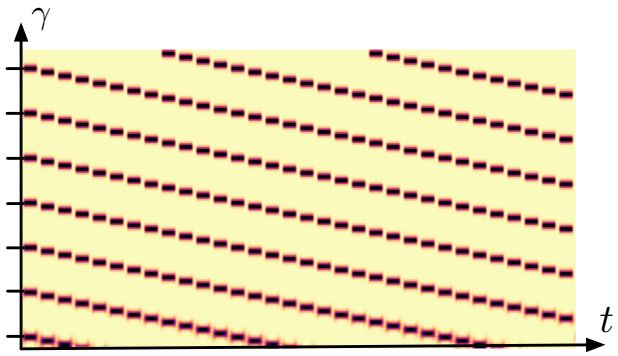


Figure 4.6: The wavelet scalogram $\mathbf{U}x(t,\gamma)$ of a chromatic scale of Shepard tones, repeated three times. The perceived pitch of $x(t)$ seems to descend endlessly.

Because $1/\sqrt{2} = \sqrt{2}/2$, the frequencies $1/\sqrt{2}$ and $\sqrt{2}$ are in octave relationship. Therefore, if the ratio between frequencies ξ' and ξ of the two Shepard tones is set exactly equal to $\sqrt{2}$, their interval is neither ascending nor descending. This mathematical edge case leads to an auditory phenomenon, called *tritone paradox*, in which the perceived direction of the interval between $x(t)$ and $x'(t)$ varies between populations (Deutsch, 1994) and subjects (Pelofi et al., 2017). We refer to Braus (1995) for a historical overview of pitch circularity and Shepard tones in Western classical music.

4.1.3 *Spatial models of pitch*

In this subsection, we address the problem of modifying the geometry of the log-frequency axis to account for octave equivalence, yet without breaking its rectilinear topology. We give a parametric equation of a three-dimensional helix in cylindrical coordinates and of a two-dimensional spiral in polar coordinates. We relate our work to the state of the art in feature engineering by constructing the so-called chroma features, which result from the application of octave equivalence onto the scalogram.

Three-dimensional helix

A way to reconcile the circularity of pitch chroma with the rectilinearity of pitch height is to twist the log-frequency axis into an ascending

helix which makes a full turn at every octave (Shepard, 1964). A parametric equation of this helix in Cartesian coordinates, expressed in function of the log-frequency variable γ, is

$$\begin{cases} x &= \cos(2\pi\gamma) \\ y &= \sin(2\pi\gamma) \\ z &= \gamma. \end{cases}$$

Figure 4.7 shows the Shepard pitch helix in perspective. Observe that the coordinates x and y draw a circle of radius 1 at an angular frequency 2π, whereas z grows monotonically with log-frequency. Because the distance to the vertical axis z is constant, it is convenient to replace Cartesian coordinates by cylindrical coordinates

$$\begin{cases} \theta &= 2\pi \times \{\gamma\} \\ z &= \gamma \\ \rho &= 1, \end{cases}$$

consisting of azimuth θ, altitude z, and constant radius ρ, and the bracket notation $\{\gamma\}$ denotes the fractional part of the number γ. In cylindrical coordinates, the variables θ and z match the notions of pitch chroma and pitch height.

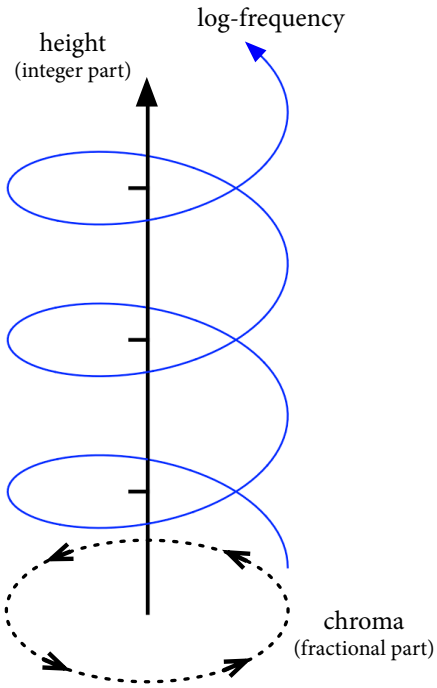


Figure 4.7: Pitch as represented on a helix. Ticks on the vertical axis denote octave intervals.

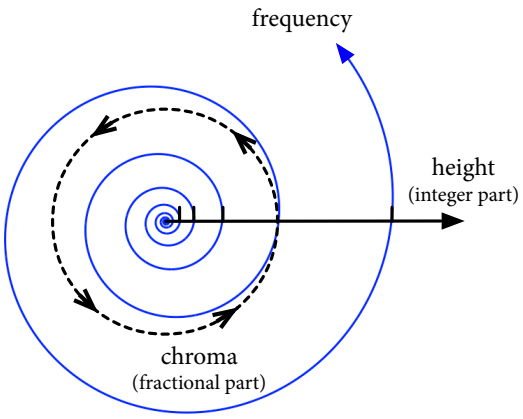


Figure 4.8: Pitch as represented on a logarithmic spiral. Ticks on the height axis denote octave intervals.

Two-dimensional spiral

It appears from the above that the pitch helix, despite being three-dimensional, can be parametrized with only two independent coordinates. For visual purposes, it can thus be replaced by a curve in two dimensions

$$\begin{cases} x &= 2^\gamma \times \cos(2\pi\gamma) \\ y &= 2^\gamma \times \sin(2\pi\gamma), \end{cases}$$

i.e. a logarithmic spiral of the frequency variable 2^γ. Again, converting the Cartesian coordinates (x, y) to polar coordinates

$$\begin{cases} \theta &= 2\pi \times \{\gamma\} \\ \rho &= 2^\gamma \end{cases}$$

provides a simpler representation of the logarithmic spiral. Like in the pitch helix, the azimuth variable θ corresponds to pitch chroma. Nevertheless, observe that the radius variable ρ, associated to pitch height, grows linearly with frequency ω, not logarithmically. This is because ρ, unlike z in the helix, must be nonnegative for polar coordinates to be well-defined. Figure 4.8 shows the logarithmic spiral of pitch.

Within a wavelet framework, we define the log-frequency variable $\gamma = \log_2 \omega$ in base two, and adopt the notation $\lfloor \gamma \rfloor$ (resp. $\{\gamma\}$) to express the integer (resp. fractional) part of γ. We decompose γ into

$$\lambda = \lfloor \lambda \rfloor + \{\lambda\} = \rho + \chi, \tag{4.3}$$

where the octave index ρ is an integer and the chroma index χ belongs to the continuous interval $[0; 1[$.

As explained in the arithmetical argument in favor of octave equivalence (end of Subsection 4.1.1), choosing an integer number Q of wavelets per octave allows to maximize overlap with harmonic structures. Within a discrete framework, $Q\gamma$ is integer, and writes as $Q\gamma = Q\rho + \chi$ where $\rho \geq 0$ is integer and χ belongs to the integer range $[\![0; Q[\![$. Therefore, the decomposition of log-frequency λ into octave ρ and chroma χ amounts to an Euclidean division of the log-frequency index $Q\gamma$ by Q, whose dividend is ρ and whose remainder is χ. When applied to some tensor

$$\mathbf{Y_2}\boldsymbol{x}(t, \gamma, \gamma_2) = (\mathbf{U_1}\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{2^{\gamma_2}})(t, \gamma)$$

of complex-valued temporal scattering coefficients, this Euclidean division reshapes $\mathbf{Y_2}\boldsymbol{x}(t, \gamma, \gamma_2)$ into a four-dimensional tensor

$$\widetilde{\mathbf{Y_2}}\boldsymbol{x}(t, \chi, \rho, \gamma_2) = \mathbf{Y}_2\boldsymbol{x}(t, \chi + Q\rho, \gamma_2)$$

under a convention of column-major array ordering. The same applies to time-frequency scattering coefficients with the necessary changes having been made.

Chroma features

Summing the responses of two wavelets separated by one octave, in order to enhance harmonic intervals in the scalogram, was pioneered by Kronland-Martinet (1988). This approach has been generalized to the full hearing range, hence leading to the so-called *chroma features*

$$\mathrm{chroma}(t, \chi) = \sum_{\rho} \mathbf{U}_1\boldsymbol{x}(t, \rho + \chi),$$

indexed by time t and pitch chroma χ. Introduced by Wakefield (1999), chroma features are widespread in music information retrieval, and automatic chord estimation in particular (McVicar et al., 2014). Other applications include audio thumbnailing (Bartsch and Wakefield, 2005), musical genre recognition (Pérez-Sancho, Rizo, and Inesta, 2009), cover song identification (Ellis, 2007), and lyrics-to-audio alignment (Mauch, Fujihara, and Goto, 2012). The number of chroma features is equal to the number of wavelets per octave Q. Observe that the chroma features refocus all the harmonic partials in a Shepard tone onto a single bump of chroma χ. Therefore, they provide a

sparser representation of Shepard tones than the wavelet scalogram, with the additional upside of a reduced dimensionality. However, the presence of non-power-of-two harmonics in real-world sounds also activates other chroma bins than χ, hence a reduced sparsity and discriminability. Consequently, most of the work on feature engineering around chroma features has been devoted to mitigating the influence of upper partials above the fundamental frequency (Cho and Bello, 2014; Jiang et al., 2011; Müller, Ewert, and Kreuzer, 2009).

Despite the loss of information brought by the sum across octaves, the interest of chroma features resides in their invariance to octave transposition. Moreover, when multiple notes overlap in time, thus forming a musical chord, chroma features are invariant to the octave transposition of the root note of chord — a common transformation in music known as chord inversion (Gold, Morgan, and Ellis, 2011, Section 37.5). For this reason, chroma features are ubiquitous in the field of automatic chord estimation. We refer to McVicar et al. (2014) for a review of the state of the art. They are also found in systems aiming at retrieving more abstract musical information, such as tonality (Peeters, 2006) or song structure (Paulus, Müller, and Klapuri, 2010), as well as in cover song identification systems (Bertin-Mahieux and Ellis, 2011; Ellis and Poliner, 2007; Serra et al., 2008).

The next section is devoted to the definition of the spiral scattering transform, which incorporates octave equivalence into a multiresolution analysis framework in the time-frequency domain.

4.2 SPIRAL SCATTERING

After rolling up the log-frequency axis on a spiral, the octave ρ appears as a discrete, radial dimension, linked which the perception of pitch height. Harmonic sounds are regular over this dimension, because their power-of-two partials are aligned on the same radius. Building a discrete filter bank of wavelets over the octave variable takes advantage of this regularity while preserving some locality of pitch height. In this section, we complement time-frequency scattering by performing wavelet convolutions across octaves ρ, in addition to time t and log-frequency γ. The resulting transform, called spiral scattering transform, fits into the framework of multivariable scattering operators introduced in Subsection 3.2.1.

To illustrate the physical meaning of the wavelet scale variables across octaves, this section gives a visualization of the spiral scattering coefficients on two kinds of synthetic signals, namely Shepard-Risset glissandos and arpeggios. We show experimentally that signal reconstruction from spiral scattering coefficients exhibits less artifacts that time-frequency scattering in the vicinity of harmonic onsets, while performing on par with time-frequency scattering in the absence of harmonic patterns.

4.2.1 *Spiral wavelets*

In this subsection, we define the spiral scattering transform. We construct a wavelet transform across octaves in the scalogram, and apply it after the first two convolutional operators involved in the computation of time-frequency scattering. Spiral scattering coefficients result from the application of three-dimensional "spiral wavelets", pointwise complex modulus, and low-pass filtering over time.

Wavelet filter bank across octaves

The spiral scattering transform is a cascade of three wavelet transforms over the wavelet scalogram $\mathbf{U}_1\boldsymbol{x}(t,\gamma)$, respectively applied over the time variable t, log-frequency variable γ, and octave variable $\rho = \lfloor \gamma \rfloor$, followed by the application of pointwise complex modulus and low-pass filtering over time. The first two wavelet transforms are also found in time-frequency scattering. The third wavelet transform can only be applied once γ has been decomposed into chroma χ and octave ρ, by appropriate array reshaping. Yet, if this reshaping were to be performed before the wavelet transform along γ, undesirable artifacts at octave boundaries would appear. Instead, we reshape γ into χ and ρ between the wavelet transforms along γ and the wavelet transform across ρ. A block diagram of the operations involved in spiral scattering is shown in Figure 4.9.

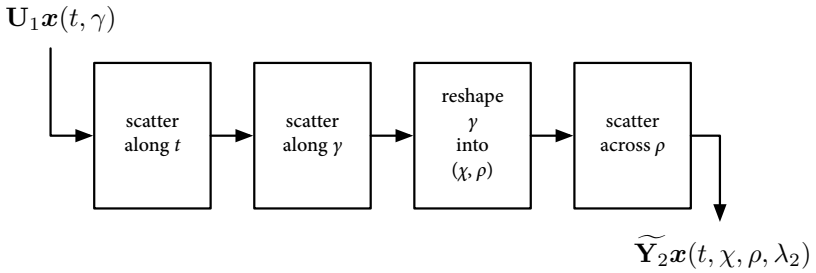


Figure 4.9: Block diagram of spiral scattering.

Because it operates over a complex input, the wavelet filter bank over the octave variable ρ must encompass negative frequencies as well as positive frequencies. It must also include a low-pass filter $\boldsymbol{\phi}$. Its definition

$$\boldsymbol{\psi}_{\theta::\rho\times 2^{\gamma::\rho}}(\rho) = \begin{cases} 2^{\gamma::\rho}\boldsymbol{\psi}((\theta::\rho)\times 2^{\gamma::\rho}) & \text{if } (\theta::\rho)\neq 0, \\ \boldsymbol{\phi}(\rho) & \text{otherwise} \end{cases}$$

is comparable to the wavelet filter bank across log-frequencies introduced in Section 3.2 about time-frequency scattering.

Spiral wavelets

The separable products of three wavelets, respectively operating over time, log-frequency and octave, turns into a multi-variable wavelet

$$\boldsymbol{\Psi}_{\lambda_2}(t,\gamma) = \boldsymbol{\psi}_{\gamma_2}(t) \times \boldsymbol{\psi}_{\theta::\gamma\times 2^{\gamma::\gamma}}(\gamma) \times \boldsymbol{\psi}_{\theta::\rho\times 2^{\gamma::\rho}}(\lfloor\gamma\rfloor)$$

over time and the Shepard pitch spiral, thereafter called a spiral wavelet. Figure 4.10 shows these three wavelets in perspective.

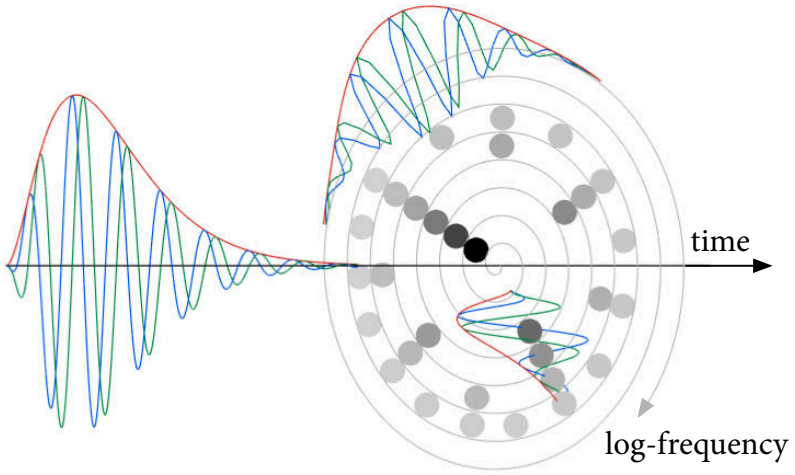


Figure 4.10: A spiral wavelet viewed in perspective. Blue and green oscillations represent the real and imaginary parts. The red envelope represents the complex modulus. Partials of a hypothetical harmonic sound are marked as thick dots.

Unrolling the Shepard pitch spiral back into a rectilinear axis allows to visualize spiral wavelets as complex-valued functions of time t and log-frequency γ. Like in Section 3.2, we adopt the shorthand notation λ_2 to denote the tuple of second-order indices

$$\lambda_2 = (\gamma_2,\, \theta::\gamma,\, \gamma::\gamma,\, \theta::\rho,\, \gamma::\rho),$$

where the *cons* operator ::, pronounced "along", has the same meaning as in Subsection 3.2.1. Figure 4.11 shows two spiral wavelets for varying values of the multi-index variable λ_2.

Spiral scattering coefficients

Convolving the scalogram with spiral wavelets leads to complex-valued spiral scattering coefficients

$$\begin{aligned}\mathbf{Y}_2\boldsymbol{x}(t,\gamma,\lambda_2) &= \mathbf{U}_1\boldsymbol{x} \overset{t,\gamma,\rho}{*} \boldsymbol{\psi}_{\lambda_2}(t,\gamma) \\ &= \mathbf{U}_1\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{\gamma_2} \overset{\gamma}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\gamma} \overset{\rho}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\rho}(t,\gamma), \qquad (4.4)\end{aligned}$$

which subsequently yield $\mathbf{U}_2\boldsymbol{x}(t,\gamma,\lambda_2)$ after pointwise complex modulus, and $\mathbf{S}_2\boldsymbol{x}(t,\gamma,\lambda_2)$ after averaging.

Figure 4.11: Spiral wavelets in the time-frequency domain. Brighter colors denote greater algebraic values of the real part. The background color corresponds to a null real part. Ticks on the vertical axis denote octave intervals.

4.2.2 *Visualization on synthetic signals*

Shepard tones are stationary signals with a precise pitch chroma but no definite pitch height — see Subsection 4.5. In this subsection, we construct two nonstationary generalizations of the Shepard tone: the Shepard-Risset glissando, which has no definite pitch height and a constant pitch chroma velocity; and the Shepard-Risset arpeggio, which has a fixed pitch chroma and a constant pitch height velocity. We draw a close correspondence between these signals and the geometry of spiral scattering coefficients. In particular, we prove that the rates of variation of pitch chroma and pitch height can be linked to the relative frequencies of spiral wavelets along time, log-frequency, and octave.

Projection from dimension five to dimension two

Let us respectively denote by ω_2, $(\omega::\gamma)$ and $(\omega::\rho)$ these center frequencies. One has $\omega_2 = 2^{\gamma_2}$, $(\omega::\gamma) = (\theta::\gamma) \times 2^{\gamma::\gamma}$, and $(\omega::\rho) = (\theta::\rho) \times 2^{\gamma::\rho}$.

Within a continuous framework, spiral scattering coefficients can be regarded as a real-valued tensor $\mathbf{U}_2\boldsymbol{x}(t, \gamma, \omega_2, \omega::\gamma, \omega::\rho)$ in dimension five. Among these five dimensions, only a subset of two can be viewed on the same graph. Therefore, in the following figures display-

ing spiral scattering coefficients, the variables t, γ, and ω_2 are fixed, while $(\omega::\gamma)$ and $(\omega::\rho)$ are plotted as horizontal and vertical axes.

Shepard-Risset glissando

A nonstationary generalization of Shepard tones can be made by the action of a time warp $\mathcal{W}_\tau \boldsymbol{x}(t) = \dot{\boldsymbol{\tau}}(t) \times (\boldsymbol{x} \circ \boldsymbol{\tau})(t)$ onto a Shepard tone $\boldsymbol{x}(t) = \sum_\rho \cos(2\pi 2^\rho \xi t)$ of pitch ξ, leading to

$$\mathcal{W}_\tau \boldsymbol{x}(t) = \dot{\boldsymbol{\tau}}(t) \times \sum_\rho 2^\rho \cos(2\pi 2^\rho \xi \boldsymbol{\tau}(t)).$$

This time warp is equivalent to a time-varying frequency transposition $\mathcal{F}_{\log_2 \dot{\boldsymbol{\tau}}(t)}$, where the variations of $\log_2 \dot{\boldsymbol{\tau}}(t)$ control the variations of the instantaneous pitch chroma. Setting $\log_2 \dot{\boldsymbol{\tau}}(t)$ to an affine function, e.g. $\dot{\boldsymbol{\tau}}(t) = 2^{\frac{t}{T}}$ as in Subsection 4.1.1, leads to the well-known Shepard-Risset glissando (Risset, 1969). Whereas the pitch chroma of a Shepard chromatic scale has unbounded derivatives at note onsets, the pitch chroma of the Shepard-Risset glissando evolves at a constant velocity

$$\frac{\mathrm{d}}{\mathrm{d}t} \log_2 \dot{\boldsymbol{\tau}}(t) = \frac{1}{T},$$

measured in octaves per second. Therefore, the Shepard-Risset glissando tiles the wavelet scalogram by octave-spaced diagonal stripes, as shown in Figure 4.12a.

Like in the Shepard chromatic scale, the perceived pitch of a Shepard-Risset glissando appears to ascend (or descend) endlessly, despite the periodicity of the underlying waveform $\boldsymbol{x}(t + T) = \boldsymbol{x}(t)$. As such, it triggers a paradox of pitch perception, whose visual analog is the barberpole illusion, in which the rotating motion of diagonal stripes on a cylinder gives the sensation of upwards or downwards translation (Wuerger and Shapley, 1996).

Shepard-Risset glissandos are found in Risset's early computer music compositions, such as *Computer Suite for Little Boy* (1968) and *Mutations* (1969). The sensation of infinite motion in pitch height can also be obtained by traditional polyphonic writing, without any electronics. The sound mass of string glissandi at the beginning of Iannis Xenakis's *Metastasis* (1955) is a typical example. György Ligeti, who had a great interest in fractals and paradoxical sensations, composed infinite chromatic scales in some of his piano studies, notably *Vertige* and *L'escalier du diable* (1988-1994). From the perspective of digital audio effects, Esqueda and Välimäki (2015) have shown that the same auditory illusion can be achieved by octave-spaced spectral notches upon a white noise background, as opposed to octave-based spectral peaks upon a silent background in the historical Shepard-Risset glissando.

It results from Subsection 2.3.1 that the scalogram of the Shepard-Risset glissando is approximately equal to

$$\begin{aligned}\mathbf{U}_1\mathcal{W}_\tau \boldsymbol{x}(t,\gamma) &\approx \mathbf{U}_1\boldsymbol{x}(\boldsymbol{\tau}(t), \gamma - \log_2 \dot{\boldsymbol{\tau}}(t)) \\ &\approx \sum_\rho |\widehat{\boldsymbol{\psi}}| \left(2^{\gamma - \frac{t}{T} + \rho}\right) \\ &\approx |\widehat{\boldsymbol{\psi}}| \left(2^{\left\{\gamma - \frac{t}{T} + \frac{1}{2}\right\} - \frac{1}{2}}\right),\end{aligned}$$

where the bracket notation denotes the fractional part. The scalogram of the Shepard-Risset glissando satisfies an octave periodicity equation:

$$\mathbf{U}_1\mathcal{W}_\tau \boldsymbol{x}(t,\gamma) = \mathbf{U}_1\mathcal{W}_\tau \boldsymbol{x}(t,\gamma+1).$$

Within the geometry of the Shepard pitch spiral, octave periodicity implies that, at every instant t, $\mathbf{U}_1\mathcal{W}_\tau \boldsymbol{x}(t,\gamma)$ is constant across the radial dimension ρ for every chosen angle χ. Because wavelets across octaves $\boldsymbol{\psi}_{(\theta,\gamma)::\rho}(\rho)$ are designed to have null mean, we deduce that spiral scattering coefficients $\mathbf{U}_2\mathcal{W}_\tau \boldsymbol{x}(t,\gamma,\omega_2,\omega::\gamma,\omega::\rho)$ are nonzero only in the edge case $\omega::\rho = 0$, that is, if $\boldsymbol{\psi}_{(\theta,\gamma)::\rho}(\rho)$ is not a wavelet but a low-pass filter $\boldsymbol{\phi}(\rho)$ across octaves.

Moreover, because the scalogram $\mathbf{U}_1\mathcal{W}_\tau \boldsymbol{x}(t,\gamma)$ consists of octave-spaced oblique stripes of temporal period T and spectral period 1, its two-dimensional gradient

$$\nabla \mathbf{U}_1\mathcal{W}_\tau \boldsymbol{x}(t,\gamma) = \begin{pmatrix} \frac{\partial}{\partial t}\mathbf{U}_1\mathcal{W}_\tau \boldsymbol{x}(t,\gamma) \\ \\ \frac{\partial}{\partial \gamma}\mathbf{U}_1\mathcal{W}_\tau \boldsymbol{x}(t,\gamma) \end{pmatrix}$$

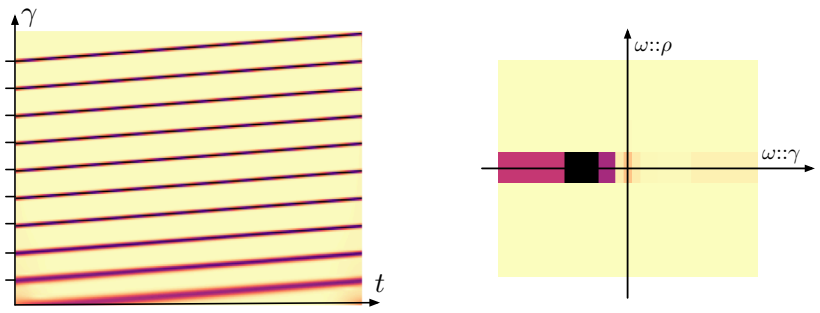
in the time-frequency domain is orthogonal to the vector $\binom{T}{1}$ over the support of the stripes, and almost zero elsewhere. Because wavelets $\boldsymbol{\psi}_{\gamma_2}(t)$ and $\boldsymbol{\psi}_{(\theta,\gamma)::\gamma}(\gamma)$ jointly respond to oriented edges in the time-frequency plane, and that the orientation of these edges is given by the gradient $\nabla \mathbf{U}_1\mathcal{W}_\tau \boldsymbol{x}(t,\gamma)$, we conclude that the amplitude of spiral scattering coefficients is maximal for frequency tuples $(\omega_2,\omega::\gamma,\omega::\rho)$ satisfying

$$\omega::\gamma = -\omega_2 T \quad \text{and} \quad \omega::\rho = 0.$$

Figure 4.12b displays $\mathbf{U}_2\mathcal{W}_\tau \boldsymbol{x}(t,\gamma,\omega_2,\omega::\gamma,\omega::\rho)$ for fixed t and γ, and γ_2 set to about $\log_2 \frac{3T}{2}$, that is $\omega_2 = \frac{2}{3T}$. We verify graphically the presence of a bump at $\omega::\gamma = -\frac{2}{3}$ and $\omega::\rho = 0$, thus confirming the theory.

Shepard-Risset arpeggio

Shepard tones with a varying pitch height, obtained by means of a time-varying broadband filter, can be visualized with spiral scattering



(a) Wavelet scalogram. (b) Spiral scattering coefficients.

Figure 4.12: The Shepard-Risset glissando: (a) time-frequency domain, (b) spiral scattering domain. In the spiral scattering domain, the horizontal (resp. vertical) axis corresponds to frequencies along neighboring log-frequencies (resp. across octaves). We observe a bump over the horizontal axis.

in a comparable way as Shepard-Risset glissandos. Although Risset has used them deliberately in some of his compositions, notably *Invisible Irène* (1995) for soprano and electronics, these sounds do not have a commonly acknowledged name. To address this lack of terminology, we suggest calling them Shepard-Risset arpeggios.

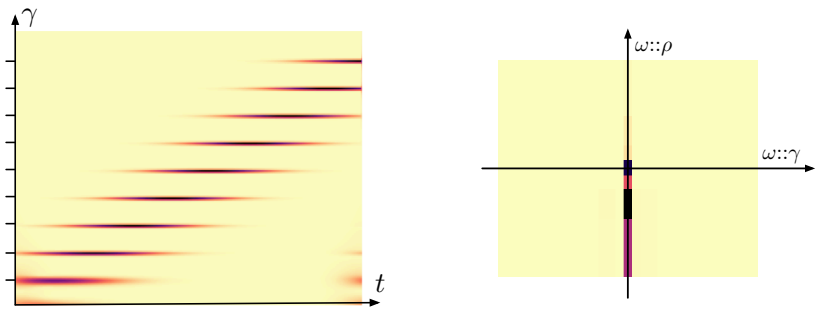
A Shepard-Risset arpeggio is a convolution between a Shepard tone $\boldsymbol{x}(t) = \sum_\rho 2^\rho \cos(2\pi 2^\rho \xi t)$ and a time-varying filter $\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h}(t)$ of relatively large bandwidth. In the sequel, we define $\boldsymbol{h}(t)$ as a Gabor time-frequency atom of quality factor equal to 1, that is, of bandwidth equal to one octave. Like in the glissando, we set $\dot{\boldsymbol{\eta}}(t) = 2^{\frac{t}{T}}$, hence inducing a constant velocity of pitch height along the log-frequency axis. In the asymptotic regime $2^\gamma Q \ll T$, the spectral envelopes of the Shepard tone and the time-varying filter get factorized in the wavelet scalogram:

$$\mathbf{U}_1(\boldsymbol{x} * \mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t,\gamma) \approx |\widehat{\boldsymbol{\psi}}|\left(2^{\left\{\gamma+\frac{1}{2}\right\}-\frac{1}{2}}\right) \times |\widehat{\boldsymbol{h}}|\left(2^{\gamma-\frac{t}{T}}\right).$$

The scalogram of a Shepard-Risset arpeggio, shown in Figure 4.13a, consists of disjoint horizontal segments arranged in a staircase-like pattern. The two-dimensional gradient of $\mathbf{U}(\boldsymbol{x} * \mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t,\gamma)$ in the vicinity of a segment is orthogonal to the vector $\binom{1}{0}$. Consequently, spiral scattering coefficients $\mathbf{U}_2(\boldsymbol{x} * \mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t,\gamma,\omega_2,\omega{::}\gamma,\omega{::}j)$ are maximal if the orientation of the time-frequency wavelet $\boldsymbol{\psi}_{\gamma_2}(t) \times \boldsymbol{\psi}_{(\theta,\gamma){::}\gamma}(\gamma)$ is vertical, that if $\omega{::}\gamma = 0$. Furthermore, the time-octave oblique periodicity

$$\mathbf{U}(\boldsymbol{x} * \mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t,\gamma) = \mathbf{U}(\boldsymbol{x} * \mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t+T,\gamma+1)$$

reveals a two-dimensional progressive wave over the variables of time t and octave ρ, whose orientation is orthogonal to the vector $\binom{T}{-1}$. We



(a) Wavelet scalogram. (b) Spiral scattering coefficients.

Figure 4.13: Shepard-Risset arpeggio: (a) time-frequency domain, (b) spiral scattering domain. In the spiral scattering domain, the horizontal (resp. vertical) axis corresponds to frequencies along neighboring log-frequencies (resp. across octaves). We observe a bump over the vertical axis.

conclude that the amplitude of spiral scattering coefficients is maximal for frequency tuples $(\omega_2, \omega::\gamma, \omega::j)$ satisfying

$$\omega::\gamma = 0 \quad \text{and} \quad \omega::\rho = -\omega_2 T.$$

Figure 4.13b displays $\mathbf{U}_2(\boldsymbol{x} * \mathcal{W}_\eta \boldsymbol{h})\boldsymbol{x}(t, \gamma, \omega_2, \omega::\gamma, \omega::j)$ for fixed t and γ, and γ_2 set to about $\log_2 3T$, that is $\omega_2 = \frac{1}{3T}$. We verify graphically the presence of a bump at $\omega::\gamma = 0$ and $\omega::j = -\frac{1}{3}$, thus confirming the theory.

Conclusion

At this stage, it appears that the Shepard-Risset glissandos and arpeggios are somewhat dual of each other, under the light of spiral scattering. Indeed, Shepard-Risset glissandos (resp. arpeggios) exhibit variations in pitch chroma (resp. height) at constant speed, while keeping pitch height (resp. chroma) constant through time. As such, their nonzero spiral scattering coefficients are concentrated around a bump, located over the horizontal (resp. vertical) axis of wavelet frequencies $\omega::\gamma$ along log-frequencies (resp. $\omega::\rho$ across octaves). Although they are unrealistic models for natural sounds, their definition helps to grasp the meaning of both axes $\omega::\gamma$ and $\omega::\rho$ in spiral scattering, along with a theoretical interpretation.

It remains to be understood how the spiral scattering coefficients behave in more realistic sound models, and under what assumptions. The next section is devoted to the study of the spiral scattering coefficients of time-varying harmonic sounds.

4.3 NONSTATIONARY SOURCE-FILTER MODEL

The source-filter model of sound production, introduced in Subsection 2.4.1, convolves a harmonic glottal source $e(t)$ with a formantic filter $h(t)$. Because this model is stationary, it cannot realistically account for variations in amplitude, pitch, and spectral envelope throughout transient regions of sound.

This section introduces a generalization of the source-filter model, by means of smooth time warps $\tau(t)$ and $\eta(t)$ applied respectively to the source and the filter, and a multiplicative amplitude $\alpha(t)$. We provide a partial differential equation satisfied by the wavelet scalogram of the nonstationary source-filter model. Furthermore, under assumptions of harmonicity and spectral smoothness, we prove that spiral scattering stably disentangles the effects of $\alpha(t)$, $\theta(t)$ and $\eta(t)$, and extracts their first and second derivatives. We conclude with the visualization of spiral scattering coefficients in synthetic source-filter signals as well as realistic instrumental sounds.

4.3.1 *Source-filter time warps*

The main drawback of the stationary source-filter model is that it is only realistic over a very short time scale. Indeed, the gestural dynamics of musical interpretation induce spectrotemporal deformations, which are specific to the instrument, at the time scale of a full musical note. For instance, the attack and release parts of a musical note are often not amenable to the same stationary model (Essid et al., 2005). Furthermore, musical articulation tends to obliterate the temporal boundaries between consecutive notes, as it replaces isolated stationary events by one nonstationary stream with smooth deformations. The same arises in speech, in which consecutive phonemes tend to be co-articulated by the speaker, that is, to get merged into a continuous sound stream without clear phonetic boundaries.

In order to reach a longer time scale in the modeling of harmonic sounds, we propose to introduce three time-varying factors in the stationary source-filter model. These factors are amplitude modulation $\alpha(t)$, source warp $\tau(t)$, and filter warp $\eta(t)$, in the fashion of what was presented in Subsections 2.3.1, 2.3.3, and 4.2.2. The resulting class of nonstationary signals offers more flexibility that the stationary model while remaining mathematically tractable.

Warped source

First, the warped source is expressed as

$$\begin{aligned}(\mathcal{W}_\tau e)(t) &= \dot{\tau}(t) \times (e \circ \tau)(t) \\ &= \dot{\tau}(t) \times \sum_{p=1}^{P} \cos(2\pi p \tau(t)).\end{aligned}$$

Our goal is to represent $(\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{e})(t)$ in the time-frequency domain so that the effect of source warp $\mathcal{W}_{\boldsymbol{\tau}}$ is approximately converted into a translating motion over the log-frequency axis. To do so, an appropriate constant-Q filter bank of wavelets should comply with the two following asymptotic conditions:

$$\frac{\ddot{\boldsymbol{\theta}}(t)}{\dot{\boldsymbol{\theta}}(t)^2} \ll \frac{1}{Q} \quad \text{and} \quad P < Q.$$

The first inequality asserts that the frequency $\dot{\boldsymbol{\theta}}(t)$ has small relative variations over the temporal support $2^{\gamma}Q$ of the corresponding wavelet $\boldsymbol{\psi}_{2^{\gamma},T(2^{\gamma})}(t)$ for $\gamma = \log_2 \dot{\boldsymbol{\theta}}(t)$, thus making partial tracking possible. Conversely, the second inequality asserts that the spectral resolution of wavelets is sufficient to segregate the P^{th} partial from its neighbor. More precisely, it means that the bandwidth $2^{\gamma}/Q$ of the wavelet $\boldsymbol{\psi}_{2^{\gamma},T(2^{\gamma})}(t)$ for $\gamma = \log_2 P + \log_2 \dot{\boldsymbol{\theta}}(t)$ should be lower than the frequency interval $P\dot{\boldsymbol{\theta}}(t) - (P-1)\dot{\boldsymbol{\theta}}(t) = \dot{\boldsymbol{\theta}}(t)$. In our numerical experiments, we set $P = 8$, $Q = 16$, $\dot{\boldsymbol{\theta}}(t) = 140$ Hz and $\ddot{\boldsymbol{\theta}}(t)/\dot{\boldsymbol{\theta}}(t) = 0.78$ c/o. Thus, the first inequality is satisfied by a factor of ten, while the second inequality is satisfied by a factor of two.

In the asymptotic regime, the wavelet ridge theorem of Delprat et al. (1992) — see Subsection 2.2.3 — may be readily applied to each nonstationary partial, yielding

$$\begin{aligned} \mathbf{U}_1(\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{e})(t,\gamma) &\approx \dot{\boldsymbol{\tau}}(t) \times \sum_{p=1}^{P} \left| \widehat{\boldsymbol{\psi}}_{2^{\gamma},T(2^{\gamma})}\left(p\dot{\boldsymbol{\tau}}(t)\right) \right| \\ &= \mathbf{U}_1\boldsymbol{e}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)). \end{aligned} \tag{4.5}$$

Warped filter

Secondly, we express the warped filter as $(\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t) = \dot{\boldsymbol{\eta}}(t) \times (\boldsymbol{h} \circ \boldsymbol{\eta})(t)$. The computation of $\mathbf{U}_1(\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t,\gamma)$ relies on the assumption that the Fourier spectrum $\left|\widehat{\boldsymbol{h}}(\omega)\right|$ of the filter has small relative variations over the support of the Fourier transform of the wavelet $\boldsymbol{\psi}_{2^{\gamma},T(2^{\gamma})}(\omega)$:

$$\frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)^2} \ll \frac{1}{Q} \quad \text{and} \quad \frac{\mathrm{d}(\log|\widehat{\boldsymbol{h}}|)}{\mathrm{d}\omega}(2^{\gamma}) \ll 2^{-\gamma}Q$$

Therefore, the convolution between $(\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t)$ and $\boldsymbol{\psi}_{\lambda}(t)$ can be approximately factorized as

$$\left((\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h}) * \boldsymbol{\psi}_{\gamma}\right)(t) \approx \boldsymbol{\psi}_{\gamma}\left(\frac{\boldsymbol{\eta}(t)}{\dot{\boldsymbol{\eta}}(t)}\right) \times \widehat{\boldsymbol{h}}\left(\frac{2^{\gamma}}{\dot{\boldsymbol{\eta}}(t)}\right). \tag{4.6}$$

An upper bound for the approximation error in the above factorization is given in Appendix B. Interestingly, this result is a swapped formulation of the wavelet ridge theorem of Delprat et al. (1992).

Whereas the wavelet ridge theorem assumes that the signal has slowly varying amplitude and instantaneous frequency over the temporal support of the wavelet, here we assume that the wavelet has slowly varying amplitude and instantaneous frequency over the temporal support of the signal. This duality between analyzed signal and analyzing wavelet is emphasized in the textbook of Flandrin (1998, Section 3.2).

From the previous equation, we deduce that the scalogram of the deformed filter $(\mathcal{W}_\eta \boldsymbol{h})(t)$ can be obtained by translating the stationary wavelet spectrum $\mathbf{U}_1 \boldsymbol{h}(\gamma)$ along the log-frequency axis γ, with a trajectory governed by $\log_2 \dot{\boldsymbol{\eta}}(t)$, hence the equation

$$\mathbf{U}_1(\mathcal{W}_\eta \boldsymbol{h})(t, \gamma) = \mathbf{U}_1 \boldsymbol{h}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t)) \tag{4.7}$$

which is comparable to the Equation 4.5 yielding the scalogram of the deformed source.

Amplitude modulation

Thirdly, the asymptotic condition

$$\frac{\dot{\boldsymbol{\alpha}}(t)}{\boldsymbol{\alpha}(t)} \ll \frac{2^\gamma}{Q}$$

guarantees that the amplitude $\boldsymbol{\alpha}(t)$ has slow modulations over the temporal support $2^\gamma Q$ of the analyzing wavelet $\boldsymbol{\psi}_{2^\gamma, T(2^\gamma)}(t)$ at the frequency 2^γ, for every γ of interest. It follows that $\boldsymbol{\alpha}(t)$ can be factorized in front of the scalogram of the nonstationary source-filter model.

Full model

Because the impulse responses of the source and the filter remain constant through time, the stationary source-filter model $\boldsymbol{x}(t) = (\boldsymbol{e} * \boldsymbol{h})(t)$ can be expressed as a convolution. If the source $\boldsymbol{e}(t)$ is deformed by a time warp $\mathcal{W}_\tau$, the resulting nonstationary source-filter model $\boldsymbol{x}(t) = \boldsymbol{\alpha}(t) \times (\mathcal{W}_\tau \boldsymbol{e} * \boldsymbol{h})(t)$ can still be expressed as a convolution, because the filtering with $\boldsymbol{h}(t)$ is a linear, translation-covariant operator. By the wavelet ridge theorem, its scalogram is equal to

$$\begin{aligned} \mathbf{U}_1 \boldsymbol{x}(t, \gamma) &\approx \boldsymbol{\alpha}(t) \times \mathbf{U}_1 \boldsymbol{e}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)) \times |\hat{\boldsymbol{h}}(2^\gamma)| \\ &\approx \boldsymbol{\alpha}(t) \times \mathbf{U}_1 \boldsymbol{e}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)) \times \mathbf{U}_1 \boldsymbol{h}(\gamma) \end{aligned}$$

Conversely, if $\boldsymbol{h}(t)$ is deformed by $\mathcal{W}_\eta$ but $\boldsymbol{e}(t)$ is not deformed, the nonstationary source-filter model $\boldsymbol{x}(t) = \boldsymbol{\alpha}(t) \times (\boldsymbol{e} * \mathcal{W}_\eta \boldsymbol{h})(t)$ can still be expressed as a convolution, whose scalogram is

$$\mathbf{U}_1 \boldsymbol{x}(t, \gamma) \approx \boldsymbol{\alpha}(t) \times \mathbf{U}_1 \boldsymbol{e}(\gamma) \times \mathbf{U}_1 \boldsymbol{h}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t)).$$

However, if both the source and the filter are deformed, the corresponding signal $\boldsymbol{x}(t)$ cannot be expressed as a convolution. Instead,

the action of the nonstationary filter should be expressed as a pseudo-differential operator applied onto the deformed source $(\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{e})(t)$. Pseudo-differential operators are beyond the scope of this dissertation, so we do not provide a closed form expression for $\boldsymbol{x}(t)$ in the general nonstationary case. Fortunately, the effects of source and filter are disentangled and factorized in the time-frequency domain:

$$\begin{aligned}\mathbf{U}_1\boldsymbol{x}(t,\gamma) &\approx \boldsymbol{\alpha}(t) \times \mathbf{U}_1(\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{e})(t,\gamma) \times \mathbf{U}_1(\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t,\gamma) \\ &\approx \boldsymbol{\alpha}(t) \times \mathbf{U}_1\boldsymbol{e}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)) \times \mathbf{U}_1\boldsymbol{h}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t)).\end{aligned} \tag{4.8}$$

Taking the logarithm of both sides converts the above product into the sum

$$\begin{aligned}\log \mathbf{U}_1\boldsymbol{x}(t,\gamma) &\approx \log_2 \boldsymbol{\alpha}(t) \\ &+ \log \mathbf{U}_1\boldsymbol{e}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)) \\ &+ \log \mathbf{U}_1\boldsymbol{h}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t)).\end{aligned} \tag{4.9}$$

Unlike in Equation 2.19, the scalograms of the filter and the source do not solely depend on γ, but also on t. Yet, the most common feature engineering techniques based on time-frequency representations, such as differentiation, cosine transform, peak-picking, or smoothing, operate over either of these variables, but not both.

In order to disentangle the contributions of the source and the filter in the equation above, it is necessary to process t and γ jointly, and capitalize on the fact that the velocities of $\boldsymbol{\alpha}(t)$, $\log_2 \dot{\boldsymbol{\tau}}(t)$, and $\log_2 \dot{\boldsymbol{\eta}}(t)$ arise at different physical scales in time and frequency. More precisely, $\log_2 \dot{\boldsymbol{\tau}}(t)$ predominantly affects pitch chroma whereas $\log_2 \dot{\boldsymbol{\eta}}(t)$ predominantly affects pitch height. As it will be shown in the sequel, rolling up the log-frequency variable γ into a Shepard pitch spiral, thus revealing the variables χ and ρ, is a powerful method to segregate source deformations from filter deformations in quasi-harmonic, nonstationary signals.

4.3.2 *Optical flow equation*

Equation 4.9 expresses the log-scalogram $\log \mathbf{U}_1\boldsymbol{x}(t,\gamma)$ of the nonstationary source-filter model as a sum of three terms: the log-amplitude $\log \boldsymbol{\alpha}(t)$, the scalogram of the deformed source $\log \mathbf{U}_1\boldsymbol{e}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t))$, and the log-scalogram of the deformed filter $\log \mathbf{U}_1\boldsymbol{h}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t))$. In this subsection, we derive a partial differential equation satisfied by $\log \mathbf{U}_1\boldsymbol{x}(t,\gamma)$ along the Shepard pitch spiral.

An analogy with motion estimation

By drawing an analogy with computer vision, $\log \mathbf{U}_1\boldsymbol{x}$ can be interpreted as a one-dimensional scene, in which the log-frequency vari-

able γ is akin to a spatial variable. Extending the metaphor, $\log \boldsymbol{\alpha}(t)$ is a time-varying amount of illumination, $\log \mathbf{U}_1 \boldsymbol{e}(\gamma)$ is a sharp object and $\log \mathbf{U}_1 \boldsymbol{h}(\gamma)$ is the visual field of the observer. Furthermore, the object $\log \mathbf{U}_1 \boldsymbol{e}(\gamma)$ and the observer $\log \mathbf{U}_1 \boldsymbol{h}(\gamma)$ are in rigid motion through time, and their spatial trajectories are respectively governed by $\log_2 \dot{\boldsymbol{\tau}}(t)$ and $\log_2 \dot{\boldsymbol{\eta}}(t)$. Many time-varying parameters interact in the scene, yet the visual cortex is able to combine spatiotemporal cues to distinguish them, and grasp a full understanding of all positions and velocities. The visual metaphor presented here is the source-filter equivalent of what Omer and Torrésani (2016) introduced to estimate the velocity of a deformed stationary process.

Delineating objects in videos by clustering similar velocities of neighboring edges is a classical method in computer vision. However, these velocities are not directly available, and should be measured in small spatiotemporal neighborhoods. The optical flow equation (Fleet and Weiss, 2006) states that, near a moving edge, the inner product between the measured gradient and the motion vector is proportional to the measured temporal derivative. It is thus a linear partial differential equation (PDE), which yields a closed form result for one rigid object in dimension one, but is under-constrained if multiple moving objects overlap or if the spatial dimension is greater than one.

In the case of the nonstationary source-filter model, the spatial variable γ is one-dimensional. Yet, source and filter are two independent moving objects, so the optical flow equation is under-constrained. The situation is different in the Shepard pitch spiral. Indeed, the one-dimensional spatial variable γ is replaced by two variables χ and ρ, over which we can obtain two independent measurements of partial derivative. Moreover, the apparent motions of the object $\log \mathbf{U}_1 \boldsymbol{e}(\gamma)$ (resp. $\log \mathbf{U}_1 \boldsymbol{h}(\gamma)$) is cancelled once projected onto the variable ρ (resp. χ), because of a harmonicity (resp. spectral smoothness) property. Combining harmonicity with spectral smoothness is already at the core of state-of-the-art methods in multi-pitch estimation (Emiya, Badeau, and David, 2010; Klapuri, 2003). Our reasoning shows that these two properties can be interpreted geometrically in the Shepard pitch spiral, and highlight their crucial importance in nonstationary sounds.

Partial derivatives along time

The first stage in the construction of the optical flow equation is the computation of partial derivatives along time for every object. By applying the chain rule for differentiation, we extract these derivatives for the log-amplitude

$$\frac{\mathrm{d}(\log \boldsymbol{\alpha})}{\mathrm{d}t}(t) = \frac{\dot{\boldsymbol{\alpha}}(t)}{\boldsymbol{\alpha}(t)},$$

the log-scalogram of the deformed source

$$\frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{e})}{\mathrm{d}t}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)) = -\frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)} \times \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{e})}{\mathrm{d}\gamma}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)),$$

and the scalogram of the deformed filter

$$\frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{h})}{\mathrm{d}t}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t)) = -\frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)} \times \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{h})}{\mathrm{d}\gamma}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t)).$$

Combining the three equations above yields

$$\begin{aligned}\frac{\partial(\log \mathbf{U}_1 \boldsymbol{x})}{\partial t}(t,\gamma) \;\approx\; & \frac{\dot{\boldsymbol{\alpha}}(t)}{\boldsymbol{\alpha}(t)} \\ & - \frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)} \times \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{e})}{\mathrm{d}\gamma}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)) \\ & - \frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)} \times \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{h})}{\mathrm{d}\gamma}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t)). \quad (4.10)\end{aligned}$$

Spectral smoothness principle

The second step consists in computing spectral derivatives. Because the log-amplitude $\log \boldsymbol{\alpha}(t)$ is independent from acoustic frequency, its partial derivative along γ is equal to zero. We obtain

$$\begin{aligned}\frac{\partial(\log \mathbf{U}_1 \boldsymbol{x})}{\partial \gamma}(t,\gamma) \;=\; & \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{e})}{\mathrm{d}\gamma}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)) \\ & + \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{h})}{\mathrm{d}\gamma}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t)).\end{aligned}$$

From the two terms on the right hand side, we argue that the former dominates the latter in the vicinity of spectral peaks. The inequality

$$\frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{h})}{\mathrm{d}\gamma}(t,\gamma) \ll \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{e})}{\mathrm{d}\gamma}(t,\gamma), \quad (4.11)$$

known as the spectral smoothness principle (Klapuri, 2003), implies

$$\frac{\partial(\log \mathbf{U}_1 \boldsymbol{x})}{\partial \gamma}(t,\gamma) \approx \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{e})}{\mathrm{d}\gamma}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)),$$

which can be readily used in Equation 4.10.

Finite differences across octaves

It remains to be determined how the term $\frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{h})}{\mathrm{d}\gamma}$ could be estimated. To do so, we define the operator

$$\frac{\Delta(\log \mathbf{U}_1 \boldsymbol{e})}{\Delta \rho}(\gamma) = \log \mathbf{U}_1 \boldsymbol{e}(\gamma) - \log \mathbf{U}_1 \boldsymbol{e}(\gamma - 1),$$

which computes the difference between adjacent octaves on the pitch spiral, at constant pitch chroma. The finite difference operator applied to the discrete variable ρ plays a similar role to the aforementioned partial derivatives along t and γ. In the pitch spiral, the infinitesimal derivative $\frac{\partial}{\partial \gamma}$ measures angular contrast whereas the finite difference $\frac{\Delta}{\Delta j}$ measures radial contrast.

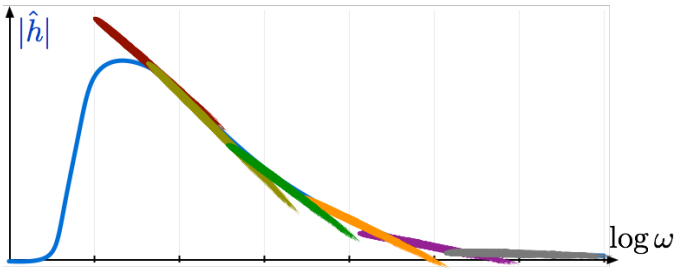


Figure 4.14: The spectral smoothness principle. The spectral envelope of a natural sound is smooth along the log-frequency axis. Consequently, it is well approximated by a piecewise linear function, up to the scale of about one octave.

Harmonicity principle

As proven in Equation 4.2, the wavelet scalogram of a Shepard tone has a periodicity of 1 in the log-frequency variable. Because the harmonic source $\boldsymbol{e}(t)$ can be represented as a sum of Shepard tones of frequencies ξ, 3ξ, 5ξ and so forth, the scalogram of $\boldsymbol{e}(t)$ satisfies

$$\mathbf{U}_1\boldsymbol{e}(\gamma) = \mathbf{U}_1\boldsymbol{e}(\gamma - 1)$$

everywhere except at odd-numbered partials, i.e. where $2^\gamma = (2p + 1)\xi$ for integer p. By analogy with computer vision, these odd-numbered partials appear as edge extremities in the pitch spiral. At a sufficiently high acoustic frequency, we may assume that the variation of contrast caused by $\mathbf{U}_1\boldsymbol{h}(\gamma)$ is more important than the variation of contrast caused by the extremity of the edge, resulting in the inequality

$$\frac{\Delta(\log \mathbf{U}_1\boldsymbol{e})}{\Delta\rho}(\gamma) \ll \frac{\Delta(\log \mathbf{U}_1\boldsymbol{h})}{\Delta\rho}(\gamma). \tag{4.12}$$

In this setting, the limit case of a Shepard tone would correspond to an infinitely long, straight edge with no extremities, hence an ideal validity of the above formula. Figure 4.15 illustrates that odd-numbered partials in a harmonic source play the role of edge extremities across the octave dimension.

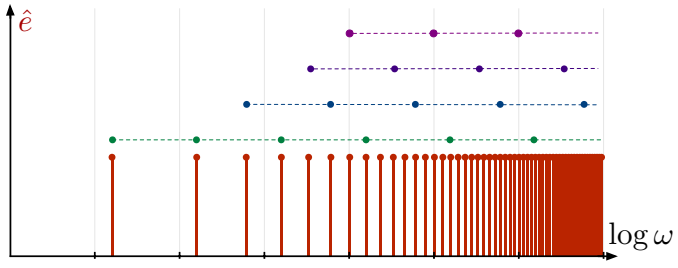


Figure 4.15: The harmonicity principle. Octave intervals in a harmonic comb draw regular edges along the octave variable, except at odd-numbered partials, which correspond to edge extremities.

The first-order Taylor expansion of $\log \mathbf{U}_1 \boldsymbol{h}(\gamma)$ allows to bound the linearization error of the approximation error of $\log \mathbf{U}_1 \boldsymbol{h}(\gamma + 1) \approx \log \mathbf{U}_1 \boldsymbol{h}(\gamma) + \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{h})}{\mathrm{d}\gamma}(\gamma)$ by

$$\left| \log \mathbf{U}_1 \boldsymbol{h}(t, \gamma + 1) - \left(\log \mathbf{U}_1 \boldsymbol{h}(t, \gamma) + \frac{\mathrm{d}(\log \mathbf{U}_1 \boldsymbol{h})}{\mathrm{d}\gamma}(t, \gamma) \right) \right| \leq \frac{1}{2} \left| \sup_{\gamma'} \frac{\mathrm{d}^2(\log \mathbf{U}_1 \boldsymbol{h})}{\mathrm{d}\gamma^2} \right|.$$

Partial differential equation

By the spectral smoothness principle, we assume that the second-order derivative of $\log \mathbf{U}\boldsymbol{x}(t, \gamma)$ along γ is small in front of the first-order derivative, an assumption which rewrites as

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} \log \left| \frac{\mathrm{d}}{\mathrm{d}\gamma} \log \mathbf{U}_1 \boldsymbol{h}(\gamma) \right| \ll 1.$$

In physics, the composition of differentiation and natural logarithm yields the relative variations of a quantity. The inequality above is constructed by cascading two of these operators, thus setting an upper bound of the relative variations of the relative variations of the log-scalogram. It should be noted that this cascade follows the same structure as the scattering transform.

Once the remainder of the Taylor expansion has been bounded, we can approximate the infinitesimal derivative $\frac{\partial(\log \mathbf{U}_1 \boldsymbol{h})}{\partial \gamma}(t, \gamma)$ by the finite difference $\frac{\Delta(\log \mathbf{U}_1 \boldsymbol{h})}{\Delta \gamma}(t, \gamma)$, which in turn is approximated by $\frac{\Delta(\log \mathbf{U}_1 \boldsymbol{x})}{\Delta \gamma}(t, \gamma)$ because of Equation 4.12. We conclude with the partial differential equation

$$\begin{aligned} \frac{\partial(\log \mathbf{U}_1 \boldsymbol{x})}{\partial t}(t, \gamma) &= \frac{\dot{\boldsymbol{\alpha}}(t)}{\boldsymbol{\alpha}(t)} \\ &- \frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)} \times \frac{\partial(\log \mathbf{U}_1 \boldsymbol{x})}{\partial \gamma}(t, \gamma) \\ &- \frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)} \times \frac{\Delta(\log \mathbf{U}_1 \boldsymbol{x})}{\Delta \rho}(t, \gamma). \end{aligned} \tag{4.13}$$

Aperture problem

At every location (t, γ) in the time-frequency plane, there are three unknowns, and yet only one optical flow equation. It is thus impossible to recover the velocities $\frac{\dot{\boldsymbol{\alpha}}(t)}{\boldsymbol{\alpha}(t)}$, $\frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)}$, and $\frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)}$ directly. This is known as the aperture problem. To solve it, it is necessary to find at least three independent equations, for various values of (t, γ), yielding the same velocities. Measuring these partial derivatives at various points in the time-frequency plane would certainly yield many equations, but they might not be independent or might relate to spurious time-frequency components.

In order to address this problem, a classical solution is to perform a multi-variable wavelet transform on $\log \mathbf{U}_1 \boldsymbol{x}(t, \gamma)$ in the vicinity of a time-frequency region (t, γ) of interest. In the context of the nonstationary source-filter model, the corresponding wavelet operator is the spiral scattering transform, as defined in section 4.2. The role of this multi-variable wavelet transform is to integrate some local context while decorrelating the contributions of multiple scales.

4.3.3 *Scattering ridge equation*

The optical flow equation of the nonstationary source-filter model, presented in Equation 4.13, exhibits an affine relationship between partial derivatives of the log-scalogram over time, log-frequency, and octaves. Furthermore, the coefficients of the equation are the velocities of amplitude, source deformation, and filter deformation. To give a more systematic description of these velocities, we integrate spectrotemporal context and decorrelate scales by applying a spiral scattering transform to the scalogram.

Recalling the definition of the spiral scattering transform — Equation 4.4 — and combining it with the approximate expression of the nonstationary source-filter model scalogram — Equation 4.8 — yields

$$\begin{aligned} \mathbf{Y}_2 \boldsymbol{x}(t, \gamma, \lambda_2) \quad &\approx \quad [\boldsymbol{\alpha}(t) \times \mathbf{U}_1 \boldsymbol{e}(\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)) \times \mathbf{U}_1 \boldsymbol{h}(\gamma - \log_2 \dot{\boldsymbol{\eta}}(t))] \\ &\quad \overset{t}{*} \boldsymbol{\psi}_{\gamma_2}(t) \overset{\gamma}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\gamma}(\gamma) \overset{j}{*} \boldsymbol{\psi}_{(\gamma,\theta)::j}(j). \end{aligned} \tag{4.14}$$

We shall prove that, for a given spectrotemporal region (t, γ) near the p^{th} partial for $p < Q$, the local maxima of $\mathbf{U}_2 \boldsymbol{x}(t, \gamma, \lambda_2) = |\mathbf{Y}_2 \boldsymbol{x}(t, \gamma, \lambda_2)|$ are clustered on a plane in the three-dimensional space $(\omega_2, (\omega::\gamma), (\omega::\rho))$ of spiral scattering coefficients. This plane satisfies the Cartesian equation

$$\frac{\dot{\boldsymbol{\alpha}}(t)}{\boldsymbol{\alpha}(t)} - \frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)} \times (\omega::\gamma) - \frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)} \times (\omega::\rho) = \omega_2. \tag{4.15}$$

Observe that this Cartesian equation follows the same structure as the optical flow equation (Equation 4.13). In the context of source separation, this result means that harmonic sounds overlapping both in time and frequency could be resolved according to their respective source-filter velocities. In the context of classification, it suggests that the spiral scattering transform is a sparse representation of harmonic sounds, even at the time scale of a full musical event comprising variations in amplitude, pitch, and timbre. Furthermore, the location of the wavelet maxima in the space $(\omega_2, (\omega::\gamma), (\omega::\rho))$ is shared among all time-frequency regions belonging to the support of $\mathbf{U}_1 \boldsymbol{x}(t, \gamma)$. The sparsity and regularity of the spiral scattering transform will be verified experimentally in Subsection 4.3.4.

Like in the previous subsection, our proof is driven by harmonicity and spectral smoothness properties. Moreover, it makes a systematic

use of the wavelet ridge theorem of Delprat et al. (1992). Indeed, this theorem is applied three times throughout the proof, i.e. once over each variable in the spiral scattering transform.

The first step consists in showing that the convolution along chromas with $\boldsymbol{\psi}_{(\gamma,\theta)::\gamma}$ only applies to the scalogram of the source $\mathbf{U}_1(\mathcal{W}_\tau \boldsymbol{e})(t,\gamma)$, whereas the convolution across octaves with $\boldsymbol{\psi}_{(\gamma,\theta)::j}$ only applies to the scalogram of the filter $\mathbf{U}_1(\mathcal{W}_\eta \boldsymbol{h})(t,\gamma)$. All wavelets are designed to have at least one vanishing moment, that is, to carry a negligible mean value. Consequently, convolving them with a constant yields zero, and convolving them with an affine function yields approximately zero. Integration by parts allows to rewrite the spectral smoothness inequality in Equation 4.11 as

$$\left|\mathbf{U}_1(\mathcal{W}_\eta \boldsymbol{h}) \overset{\gamma}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\gamma}\right| \ll \left|\mathbf{U}_1(\mathcal{W}_\tau \boldsymbol{e}) \overset{\gamma}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\gamma}\right|, \tag{4.16}$$

and the harmonicity inequality in Equation 4.12 as

$$\left|\mathbf{U}_1(\mathcal{W}_\tau \boldsymbol{e}) \overset{j}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\rho}\right| \ll \left|\mathbf{U}_1(\mathcal{W}_\eta \boldsymbol{h}) \overset{j}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\rho}\right|. \tag{4.17}$$

The amplitude term $\boldsymbol{\alpha}(t)$ is independent of acoustic frequency, so convolving it with either $\boldsymbol{\psi}_{(\gamma,\theta)::\gamma}$ or $\boldsymbol{\psi}_{(\gamma,\theta)::\rho}$ would yield zero. Henceforth, it can be factorized out from the wavelets transforms along γ and ρ. Combining the two equation above into the definition of spiral scattering in Equation 4.14 yields

$$\begin{aligned}
(\mathbf{U}\boldsymbol{x} \overset{t}{*} \boldsymbol{\psi}_{\gamma_2})(t,\gamma) \;\approx\; & \boldsymbol{\alpha}(t) \\
\times\; & \left(\mathbf{U}_1(\mathcal{W}_\tau \boldsymbol{e}) \overset{\gamma}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\gamma}\right)(t,\gamma) \\
\times\; & \left(\mathbf{U}_1(\mathcal{W}_\eta \boldsymbol{h}) \overset{\rho}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\rho}\right)(t,\gamma) \\
\overset{t}{*}\; & \boldsymbol{\psi}_{\gamma_2}(t,\gamma).
\end{aligned}$$

The next step of the proof consists in extracting the phases of the three complex-valued signal $(\boldsymbol{\alpha} \overset{t}{*} \boldsymbol{\psi}_{\gamma_2})(t)$, $(\mathbf{U}_1(\mathcal{W}_\tau \boldsymbol{e}) \overset{\gamma}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\gamma})(t,\gamma)$ and $(\mathbf{U}_1(\mathcal{W}_\eta \boldsymbol{h}) \overset{j}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\rho})(t,\gamma)$. First, we locally approximate $\boldsymbol{\alpha}(t)$ by a cosine function of amplitude a_α, fundamental frequency ξ_α and initial phase φ_α, to which is added a constant b_α:

$$\boldsymbol{\alpha}(t) \approx b_\alpha + a_\alpha \cos(2\pi\xi_\alpha t + \varphi_\alpha).$$

The equation above is a typical model for musical tremolo (Andén and Mallat, 2012). Differentiating the above through time yields $\dot{\boldsymbol{\alpha}}(t) = -2\pi\xi_\alpha a_\alpha \sin(2\pi\xi_\alpha t + \varphi_\alpha)$ and $\ddot{\boldsymbol{\alpha}}(t) = -4\pi^2\xi_\alpha^2 a_\alpha \sin(2\pi\xi_\alpha t + \varphi_\alpha)$. Going back to a nonparametric model of $\boldsymbol{\alpha}(t)$, the instantaneous frequency of $\boldsymbol{\alpha}(t)$ is locally approximated by $\frac{\ddot{\boldsymbol{\alpha}}(t)}{\dot{\boldsymbol{\alpha}}(t)}$.

Secondly, we apply the wavelet ridge theorem of Delprat et al. (1992) along γ to the term $(\mathbf{U}_1(\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{e}) \overset{\gamma}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\gamma})(t,\gamma)$. We obtain a phase of

$$(\omega::\gamma) \times (\gamma - \log_2 \dot{\boldsymbol{\tau}}(t)),$$

up to a constant phase shift. By differentiating this quantity along t for fixed γ, the instantaneous frequency of the term $(\mathbf{U}_1(\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{e}) \overset{\gamma}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\gamma})(t,\gamma)$ is equal to $-(\omega::\gamma) \times \frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)}$.

Thirdly, we apply the wavelet ridge theorem across ρ to the term $(\mathbf{U}_1(\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h}) \overset{\rho}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\rho})(t,\gamma)$. We obtain a phase of

$$(\omega::\rho) \times (\gamma - \log_2 \dot{\boldsymbol{\eta}}(t)),$$

up to a constant phase shift. Again, by differentiating this quantity, the instantaneous frequency of the term $(\mathbf{U}_1(\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h}) \overset{j}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\rho})(t,\gamma)$ is equal to $-(\omega::\rho) \times \frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)}$.

The final step of the proof consists in combining the three instantaneous frequencies obtained above, and apply the band-pass filtering with the temporal wavelet $\boldsymbol{\psi}_{2^{\gamma_2}}(t)$. As long as

$$\omega_2 \geq \left| \frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)} \times (\omega::\gamma) \right| \quad \text{and} \quad \omega_2 \geq \left| \frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)} \times (\omega::\rho) \right|,$$

the envelope of the convolutions along γ and ρ are almost constant over the support of $\boldsymbol{\psi}_{2^{\gamma_2}}(t)$. Thus, we may apply the wavelet ridge theorem a third time, now on the temporal dimension, to yield the approximate formula

$$\begin{aligned} \mathbf{U}_2\boldsymbol{x}(t,\gamma,\lambda_2) \;\approx\; & a_\alpha \\ & \times \left| \mathbf{U}_1(\mathcal{W}_{\boldsymbol{\tau}}\boldsymbol{e}) \overset{\gamma}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\gamma} \right| (t,\gamma) \\ & \times \left| \mathbf{U}_1(\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h}) \overset{\rho}{*} \boldsymbol{\psi}_{(\gamma,\theta)::\rho} \right| (t,\gamma) \\ & \times \left| \widehat{\boldsymbol{\psi}}_{2^{\gamma_2}} \left(\frac{\ddot{\boldsymbol{\alpha}}(t)}{\dot{\boldsymbol{\alpha}}(t)} - \frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)}(\omega::\gamma) - \frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)}(\omega::\rho) \right) \right| \end{aligned} \tag{4.18}$$

The Fourier spectrum $|\widehat{\boldsymbol{\psi}}_{2^{\gamma_2}}(\omega)|$ of $\boldsymbol{\psi}_{2^{\gamma_2}}(t)$ is a smooth bump centered at the modulation frequency $\omega_2 = 2^{\gamma_2}$. Equation 4.15 thus follows immediately.

4.3.4 *Experimental validation*

The reasoning presented above relies on many assumptions: asymptoticity, harmonicity, spectral smoothness, and slow modulations. In addition, the time-frequency region (t,γ) must correspond to the p^{th} partial of the source, with p not too large to avoid interferences with

neighboring partials, and also p even to avoid edge artifacts at odd-numbered partials. It begs the question whether the formula in Equation 4.15, despite its relative simplicity, is useful at all.

In order to assess whether the theoretical result of Equation 4.15 is experimentally reproducible, we generate a signal according to the nonstationary source-filter model. For the sake of simplicity, the amplitude $\boldsymbol{\alpha}(t)$ is a Planck taper window of duration equal to two seconds. Consequently, in the vicinity of the time origin $t = 0$, $\boldsymbol{\alpha}(t)$ may be replaced by a constant, and thus the term $\frac{\dot{\boldsymbol{\alpha}}(t)}{\boldsymbol{\alpha}(t)}$ disappears from the problem.

The source $\boldsymbol{e}(t)$ has a fundamental frequency of $\xi_e = 140\,\text{Hz}$, and contains $P = 8$ harmonic partials of equal amplitude. The temporal derivative $\dot{\boldsymbol{\tau}}(t)$ of the source deformation $\boldsymbol{\theta}(t)$ is of the form

$$\dot{\boldsymbol{\tau}}(t) = 2^{\frac{t}{T_\tau}},$$

where the chirp rate $\frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)} = \frac{1}{T_\tau}$ is of the order of 1 octave per second.

The filter $\boldsymbol{h}(t)$ is a Morlet wavelet of center frequency $\xi_h = 500\,\text{Hz}$ and of quality factor $Q_h = 1$. Like $\boldsymbol{\tau}(t)$, the deformation of the filter $\boldsymbol{\eta}(t)$ has a temporal derivative of the form

$$\dot{\boldsymbol{\eta}}(t) = 2^{\frac{t}{T_\eta}},$$

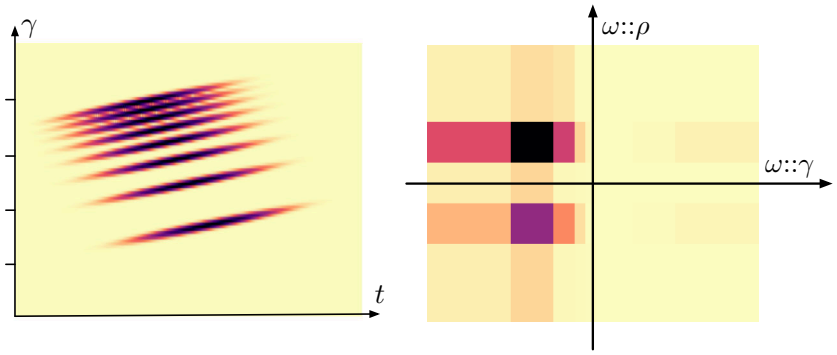
where the rate $\frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)} = \frac{1}{T_\eta}$ is of the order of -4 octaves per second. The audible result, $\boldsymbol{x}(t)$, is an upwards glissando with a timbre getting progressively warmer as the passband of the nonstationary filter reaches lower frequencies.

The signal $\boldsymbol{x}(t)$ is analyzed by an Morlet filter bank of quality factor $Q = 24$, and subsequently transformed with spiral scattering. The results are presented on Figure 4.16. For a temporal modulation frequency of $\omega_2 = 2\,\text{Hz}$, we observe a bump at $\omega{::}\gamma = -\frac{2}{3}\,\text{c/o}$ and $\omega{::}\rho = +\frac{2}{3}\,\text{c/o}$ (Figure 4.16b). We may indeed verify that

$$\begin{aligned} &\frac{\ddot{\boldsymbol{\alpha}}(t)}{\dot{\boldsymbol{\alpha}}(t)} - \frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)} \times (\omega{::}\gamma) - \frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)} \times (\omega{::}\rho) \\ = \; &0 - 1 \times \left(-\frac{2}{3}\right) - (-4) \times \left(\frac{2}{3}\right) = 2\ \text{Hz}. \end{aligned}$$

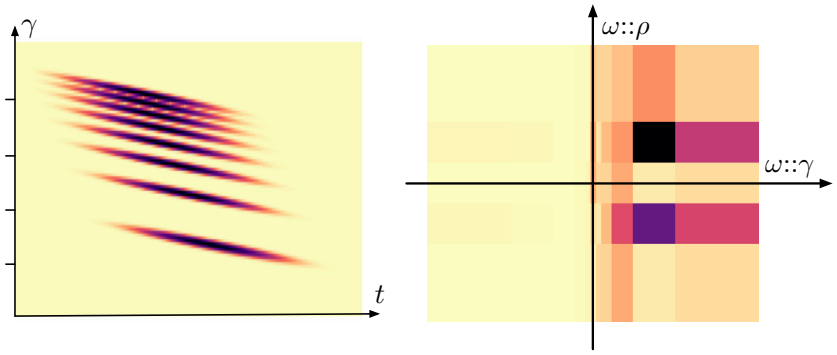
Likewise, Figures 4.16b, 4.16d, 4.16f, and 4.16h present bumps at $\omega{::}\gamma = \pm\frac{2}{3}\,\text{c/o}$ and $\omega{::}\rho = \pm\frac{2}{3}\,\text{c/o}$ depending on the sign of the ratios $\frac{\ddot{\boldsymbol{\tau}}(t)}{\dot{\boldsymbol{\tau}}(t)}$ and $\frac{\ddot{\boldsymbol{\eta}}(t)}{\dot{\boldsymbol{\eta}}(t)}$.

Despite this positive result, a secondary bump appears at the negative frequency $\omega{::}\rho = -\frac{2}{3}\,\text{c/o}$ in addition to $\omega{::}\rho = +\frac{2}{3}\,\text{c/o}$. This is because the Morlet wavelets over octaves are not rigorously analytic, as they are built upon a small, discrete array of only 8 samples. As a result, it is difficult to properly disentangle upward motion from downward motion of the nonstationary filter $(\mathcal{W}_{\boldsymbol{\eta}}\boldsymbol{h})(t)$ in the spiral scattering transform.



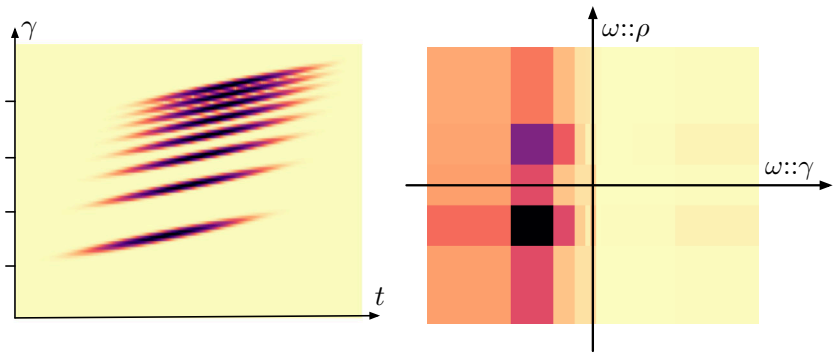
(a) Wavelet scalogram, $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)} > 0$, $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)} < 0$.

(b) Spiral scattering coefficients, $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)} > 0$, $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)} < 0$.



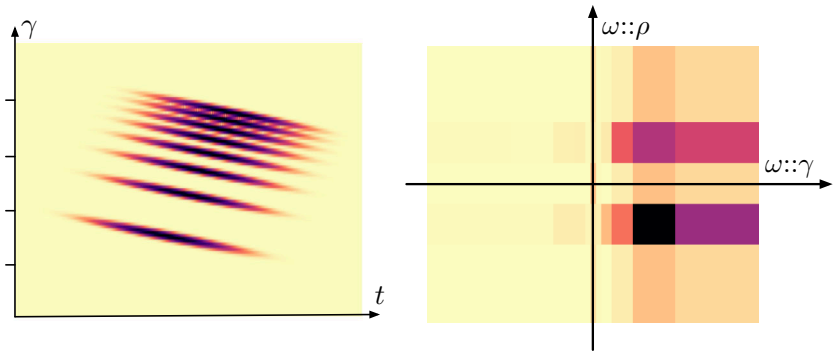
(c) Wavelet scalogram, $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)} < 0$, $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)} < 0$.

(d) Spiral scattering coefficients, $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)} < 0$, $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)} < 0$.



(e) Wavelet scalogram, $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)} > 0$, $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)} > 0$.

(f) Spiral scattering coefficients, $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)} > 0$, $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)} > 0$.



(g) Wavelet scalogram, $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)} < 0$, $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)} > 0$.

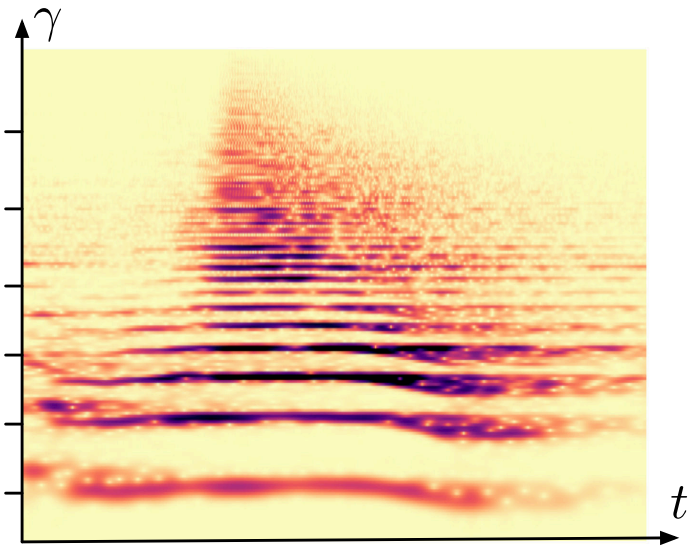
(h) Spiral scattering coefficients, $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)} < 0$, $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)} > 0$.

Figure 4.16: Spiral scattering ridges for various velocities of the source and the filter.

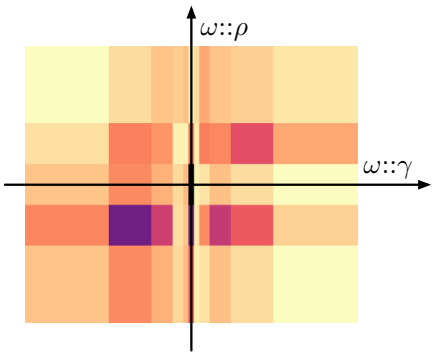
Lastly, it remains to be assessed whether the interpretation of spiral scattering coefficients as source-filter modulations could be applied to real-world sounds. With this aim in mind, we compute the spiral scattering transform of a trombone note in Berio's *Sequenza V* (1966). This piece of contemporary music contains many extended instrumental techniques, such as brassiness effects, pitch slides (*glissandi*), flutter tonguing (*flatterzunge*), and the use of a plunger mute (*sordina*). We focus on a relatively short, but complex musical note around *3'50"*, in the Stuart Dempster version from the album *The Complete Sequenzas* (2006).

The musical note, whose scalogram is shown in Figure 4.17a, can be roughly decomposed into an attack part and a release part. The attack has increasing amplitude, increasing brassiness, and increasing pitch due to the upward motion of the trombone slide. According to the nonstationary source-filter model presented above, the amplitude velocity $\frac{\dot{\alpha}(t)}{\alpha(t)}$, the source velocity $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)}$, and the filter velocity $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)}$ are all positive throughout the attack part. Conversely, the release part has decreasing amplitude, decreasing brassiness due to the use of the mute, and decreasing pitch due to the downward motion of the trombone slide. Therefore, all three velocities $\frac{\dot{\alpha}(t)}{\alpha(t)}$, $\frac{\ddot{\tau}(t)}{\dot{\tau}(t)}$, and $\frac{\ddot{\eta}(t)}{\dot{\eta}(t)}$ are supposed to be negative throughout the release part.

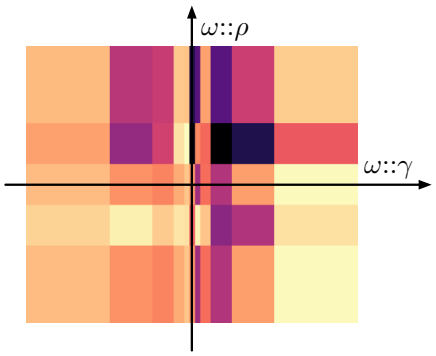
The spiral scattering coefficients are shown in Figure 4.17b for the attack part, and Figure 4.17c for the release part. The chosen time-frequency region corresponds to the fourth harmonic partial, which is in octave relationship with the fundamental. For the attack part, we observe a bump in the bottom-left quadrant, indicating an increasing pitch and an increasing brightness. For the release part, we observe a bump in the top-right quadrant, indicating a decreasing pitch and an decreasing brightness. These observations are consistent with the instrumental techniques described in the musical score. However, we also observe other activated coefficients, without theoretical explanation. Moreover, this observation only holds for a relatively restricted range of temporal modulations ω_2. In conclusion, the visualization of spiral scattering coefficients does provide some insight on the dynamics of player-instrument interaction, but remains difficult to interpret in a real-world setting.



(a) Wavelet scalogram.



(b) Spiral scattering, attack.



(c) Spiral scattering, release.

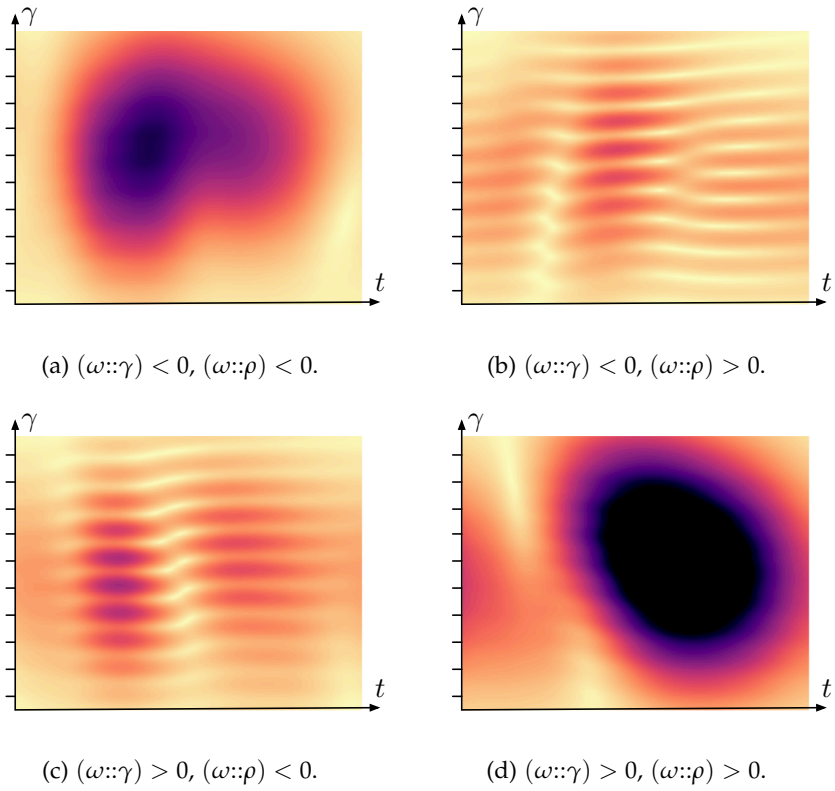
Figure 4.17: A fragment of Berio's *Sequenza V*, for trombone (1965).

A different perspective on this visualization can be obtained by fixing the value of the second-order scattering path $\lambda_2 = (\omega_2, \omega::\gamma, \omega::\rho)$, and plot the time-frequency distribution of coefficients $\mathbf{U}_2\boldsymbol{x}(t, \gamma, \lambda_2)$ as functions of t and γ. Figure 4.18 shows the corresponding heat maps for $\omega_2 = 2\,\text{Hz}$, $\omega::\gamma = \pm\frac{2}{3}\,\text{c/o}$ and $\omega::\rho = \pm\frac{2}{3}\,\text{c/o}$. The signs $\theta::\gamma = \pm 1$ and $\theta::\rho = \pm 1$ are flipped across subfigures (a), (b), (c), and (d).

In Subfigure 4.18a, we observe a large, regular bump in time and frequency near the beginning of the signal. This subfigure corresponds to negative frequencies $(\omega::\gamma)$ and $(\omega::\rho)$, i.e. the bottom-left quadrant in Subfigure 4.17b. The bump is thus associated with the attack part of the note.

Conversely, in Subfigure 4.18d, we observe a bump near the end of the signal. This subfigure corresponds to positive frequencies $(\omega::\gamma)$ and $(\omega::\rho)$, i.e. the top-right quadrant in Subfigure 4.17c. The bump is thus associated with the release part of the note.

Subfigures 4.18b and 4.18c do not present any concentrated bump of energy, because the nonstationary source-filter produces destructive interference with the corresponding spiral wavelet.



(a) $(\omega::\gamma) < 0$, $(\omega::\rho) < 0$.

(b) $(\omega::\gamma) < 0$, $(\omega::\rho) > 0$.

(c) $(\omega::\gamma) > 0$, $(\omega::\rho) < 0$.

(d) $(\omega::\gamma) > 0$, $(\omega::\rho) > 0$.

Figure 4.18: Flipping spiral wavelets.

This visualization shows that, unlike temporal scattering or time-frequency scattering, spiral scattering yields a small number of regular activations in time and frequency when applied to nonstationary harmonic sounds. Therefore, averaging locally $\mathbf{U}_2\boldsymbol{x}(t, \gamma, \lambda_2)$ into an invariant representation $\mathbf{S}_2\boldsymbol{x}(t, \gamma, \lambda_2)$ is likely to preserve most of the discriminative information for classification. Furthermore, because attack and release have been mapped to distinct scattering paths λ_2, a fully delocalized average might discard their respective locations, but will not destroy their timbral properties. Spiral scattering thus appears as an efficient way to characterize musical transients without prior detection or segmentation.

The next section presents a benchmark for the re-synthesis of locally stationary audio textures, and proceeds to show that spiral scattering is slightly better than time-frequency scattering at recovering harmonic intervals.

4.4 SIGNAL RE-SYNTHESIS FROM SPIRAL SCATTERING COEFFICIENTS

In Section 3.3, we have presented an algorithm for audio texture synthesis from averaged scattering coefficients, based on gradient backpropagation. We have shown that, for an averaging duration T of the order of 500 ms, four categories of audio textures are successfully reconstructed: decorrelated impulses (bubbling water), chirps (skylark), English speech, and broadband impulses (congas).

In this section, we address the problem of reconstructing polyphonic music, which is more complex than the aforementioned categories of sounds. We outline the steps in the gradient backpropagation of spiral scattering coefficients. We benchmark temporal scattering, time-frequency scattering, and spiral scattering in a task of audio reconstruction of a jazz excerpt. We find that spiral scattering, unlike its predecessors, is able to recover long-range interactions within the harmonic structure of pitched instruments.

4.4.1 *Gradient backpropagation of spiral scattering*

In this subsection, we describe our method for re-synthesis of audio signals from averaged spiral scattering coefficients.

The gradient backpropagation of the second-order scattering operator $\mathbf{U}_2$ towards the scalogram $\mathbf{U}_1 \boldsymbol{x}(t, \gamma)$ follows the same structure as the backpropagation of time-frequency scattering, presented in Subsection 3.3.2. except that three modulation filter banks are present instead of two. Moreover, the variables of chroma χ and octave ρ must be reshaped into log-frequency γ after backscattering complex-valued scattering coefficients $\mathbf{Y}_2 \boldsymbol{x}(t, \gamma, \lambda_2)$ across octaves. A block diagram of the operations involved in the backpropagation of spiral scattering is shown in Figure 4.19. Observe that these operations are in the reverse order with respect to spiral scattering, shown in the block diagram of Figure 4.9.

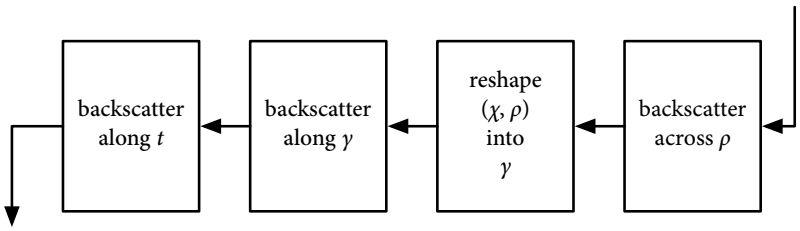


Figure 4.19: Block diagram of spiral backscattering.

4.4.2 *Experimental validation*

In this subsection, we set up a challenging benchmark of audio signal re-synthesis from averaged scattering coefficients, in which the target is a polyphonic jazz recording. We find that spiral scattering provides a slight, albeit noticeable improvement over time-frequency scattering in the reconstruction of harmonic structures.

Protocol

The protocol for audio reconstruction used in this subsection is the same as in Section 3.3: we choose a 3-second audio excerpt $\boldsymbol{x}(t)$, compute its scattering coefficients $\mathbf{S}\boldsymbol{x}(t, \lambda)$ with an averaging duration of $T = 743$ ms, and perform gradient descent to synthesize a new signal $\boldsymbol{y}(t)$ whose scattering representation $\mathbf{S}\boldsymbol{y}(t, \lambda)$ is close to $\mathbf{S}\boldsymbol{x}(t, \lambda)$.

Results on audio textures and speech

First of all, we have ran re-synthesis experiments with the four target signals presented in Subsection 3.3.3 (bubbling water, skylark, speech, and congas) and spiral scattering as the representation $\mathbf{S}\boldsymbol{x}(t, \lambda)$. We found that, in all cases, spiral scattering did not improve nor degrade the perceptual quality of the reconstructed signal with respect to time-frequency scattering.

In the case of speech, both time-frequency scattering and spiral scattering led to reconstructions that had about the same level of intelligibility. Doubling the averaging duration T to 1486 ms made the reconstructed speech unintelligible in both cases.

Results on polyphonic music

In order to evaluate the reconstruction capabilities of spiral scattering, we seek for an audio excerpt $\boldsymbol{x}(t)$ that is more challenging that the four target signals of Subsection 3.3.3. We isolate an excerpt of polyphonic jazz music from the RWC dataset (Goto et al., 2002), entitled "Lounge Away" (code name J023) and arranged for an octet comprising four saxophones.

The scalograms of all reconstructions are shown in Figure 4.20. In Figure 4.20b, we observe that reconstructing only from averaged scalogram coefficients $\mathbf{S}_1\boldsymbol{x}(t, \gamma)$ leads to a blurred result, from which transient structures have disappeared. Supplementing these coefficients with temporal scattering coefficients (Figure 4.20c) allows to recover intermittency within each subband, but does not synchronize onsets across subbands. Replacing temporal scattering by time-frequency scattering (Figure 4.20d) only leads to a minor improvement. This is because polyphonic music, unlike bioacoustic sounds, does not contain chirps. Rather, it is an intricate combination of par-

tials interfering in time, mixed with broadband structures of onsets and offsets.

The reconstruction from spiral scattering is shown in Figure 4.20e. We observe three kinds of qualitative improvements in the scalogram, with respect to time-frequency scattering. First, octave intervals in the first three notes, in which bass and drums are absent, are better synchronized. This is because spiral scattering explicitly accounts for octave equivalence. Secondly, the three bass notes are correctly retrieved and placed in time, whereas time-frequency scattering only produces two of such notes. We postulate that this is an effect of the three-fold increase in the number of coefficients, which leads to an overcomplete encoding at the lowest frequencies. Thirdly, the reverberation of the cymbal, at the end of the excerpt, is more realistic with spiral scattering than with time-frequency scattering. This is because reverberation, unlike chirps, is a broadband time-frequency pattern.

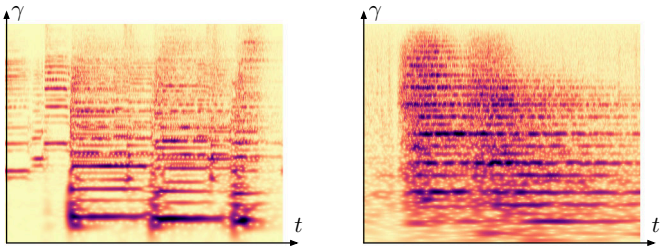
Despite the minor improvements brought by spiral scattering, many artifacts remain, especially at the highest frequencies. In particular, the interference between the cymbal onsets and the high-frequency partials is inaccurate, as it is replaced by "swooshing" effects. Moreover, the group of articulatory notes played by the horn section is missing altogether.

In conclusion, this experiment shows that faithfully recovering polyphonic music with an averaging duration T over 500 milliseconds is still out of reach of currently existing systems. The next section presents a benchmark of musical instrument classification in continuous recordings, and shows that spiral scattering outperforms time-frequency scattering and deep convolutional networks.

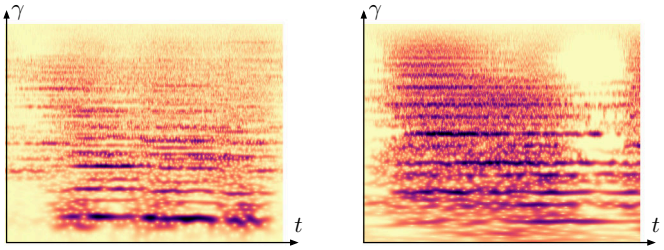
4.5 APPLICATION TO MUSICAL INSTRUMENT CLASSIFICATION

Musical performance combines a broad range of pitches, loudness dynamics, and expressive techniques. These natural factors of variability can be modeled as transformations in the time-frequency domain, which affect the spectral envelope of musical notes as well as the harmonic patterns. The challenge of musical instrument recognition amounts to building an invariant representation to these transformations while remaining discriminative to timbral properties. One possible application of this challenge is the automatic organization of ethnomusicological audio recordings (Fourer et al., 2014). More generally, it is used as a test bed for the comparison of invariant representations of musical audio (Joder, Essid, and Richard, 2009; Li, Qian, and Wang, 2015; McFee, Humphrey, and Bello, 2015).

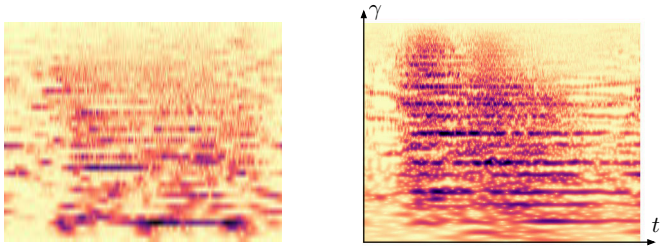
In this section, we address the problem of identifying a musical instrument from a 3-second audio recording according to a taxonomy of ten classes. Subsection 4.5.1 reviews the related work in this domain. Subsection 4.5.2 presents the chosen dataset. Subsection 4.5.3



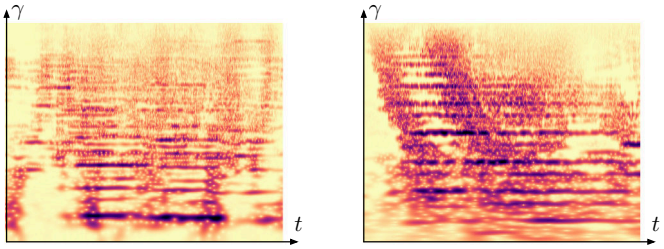
(a) Original.



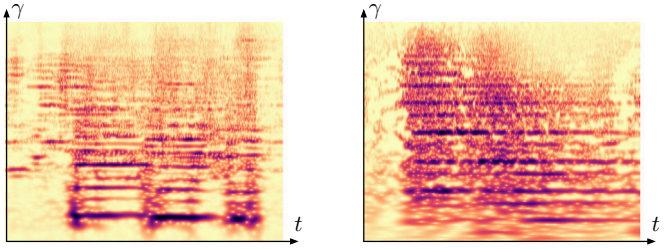
(b) Averaged scalogram only.



(c) Averaged scalogram and temporal scattering.



(d) Averaged scalogram and time-frequency scattering.



(e) Averaged scalogram and spiral scattering.

Figure 4.20: Reconstruction of a recording of polyphonic music with an averaging duration and various scattering transforms: temporal, time-frequency, and spiral. Left: jazz music, RWC-J023, $T = 743$ ms. Right: classical music, beginning of the scherzo of Beethoven's ninth symphony, $T = 500$ ms.

describes the benchmark of scattering architectures that were tried. Lastly, Subsection 4.5.4 reports state-of-the-art results with spiral scattering, thus outperforming time-frequency scattering and deep convolutional networks.

4.5.1 *Related work*

In this subsection, we review the existing publications addressing the automatic classification of musical instruments from audio. They fall into two categories, whether the data consists of isolated notes or continuous recordings. Because training a support vector machine (SVM) classifier on spectrotemporal features achieves a near-human accuracy, the problem of classifying instruments from isolated notes is now regarded as solved. However, classifying instruments from continuous recordings remains highly challenging to this day.

Automatic classification of isolated notes

Most of the datasets of monophonic instrumental sounds consist of isolated notes, played at different pitches and nuances, in well-controlled recording conditions (Eerola and Ferrer, 2008; Goto et al., 2003). Notwithstanding their usefulness for acoustics, cognitive science, and music production, these datasets are of questionable interest in the realm of content-based information retrieval. Indeed, the classification of isolated notes is deemed to be a simpler problem than the classification of a full-length musical performance, in which the articulatory effects of interpretation bring more variability to the data.

Yet, for lack of more appropriate datasets, early works on musical instrument recognition focused on isolated notes (Brown, 1999; Eronen and Klapuri, 2000; Herrera-Boyer, Peeters, and Dubnov, 2003; Kaminskyj and Czaszejko, 2005; Martin and Kim, 1998; Wieczorkowska and Żytkow, 2003). In this context, generalizing from the training set to the test set entails the construction of a representation of musical notes which is invariant to small variations in pitch and dynamics while remaining discriminative to qualitative timbre. Patil et al. (2012) have managed to classify isolated notes belonging to 11 instruments from the RWC dataset (Goto et al., 2003) with a mean accuracy of 98.7%. They used a collection of spectrotemporal receptive field (STRF) features, which are akin to time-frequency scattering features, and a support vector machine (SVM) classifier with a Gaussian kernel. Not only did they attain a near-perfect recognition rate, but they also found that the confusion matrix of the classifier was closely similar to the confusion matrix of human listeners.

Drawing on these findings, it seems that the supervised classification of musical instruments from recordings of isolated notes could now be considered a solved problem (Patil and Elhilali, 2015). Consequently, research on isolated notes is now aiming at new challenges,

such as including a broader diversity of instrumental techniques (Kostka, 2016, chapter 11) or switching to an unsupervised setting (Benetos, Kotti, and Kotropoulos, 2006). We refer to Bhalke, Rao, and Bormane (2016) for a recent review of the state of the art.

Automatic classification of continuous recordings

Continuous recordings, unlike isolated notes, remain a challenging topic for musical instrument recognition (Joder, Essid, and Richard, 2009; Krishna and Sreenivas, 2009; Livshin and Rodet, 2004; Patil and Elhilali, 2015). Owing to the variability in phrasing induced by the interpret, the required invariant for classification should encompass more sophisticated transformations than pitch shifts and changes in nuance. Moreover, it is more difficult for an automatic system to match human performance over continuous recordings than on isolated notes, because the proportion of errors decrease with the duration of the recordings, reaching virtually zero for well-trained listeners (Robinson and Patterson, 1995). The importance of phrasing in the discriminability of instruments recordings appeals for an appropriate representation of temporal modulation, to which the scattering transform is a natural candidate.

Datasets of monophonic instrumental recordings are difficult to design, because most of the commercially available music is polyphonic. It appears that the Western tradition of solo music is essentially restricted to a narrow scope of instruments (piano, classical guitar, violin), and genres (sonatas, contemporary, free jazz, folk). Without a careful curation of the musical data, undesirable correlations between instrument and genre appear. Therefore, it is no longer certain that the task at hand is genuinely about retrieving musical instruments (Sturm, 2014a).

In order to evaluate the performance of STRFs in the classification of continuous recordings, Patil and Elhilali (2015) built a dataset of 20 albums from six musical instruments, amounting to about 12 hours of audio, which was not released publicly. After performing a random split between training set and test set, they reported an average accuracy of 97.7%, which is excellent in comparison with the mere 80.0% achieved by MPEG-7 features (Casey, 2001). Nevertheless, if the split between training set and test set was performed over entire albums, not songs, the average accuracy of STRFs dropped to 80.0%, in comparison with 66% for MPEG-7 features. In other words, the task of musical instrument recognition had been implicitly made easier by collapsing the instrument labels with other sources of variability, such as artistic interpretation and sound engineering. This discrepancy in accuracies shows the crucial importance of designing a strict and reproducible split of the data into folds, in order to provide a fair evaluation of machine learning algorithms (Sturm, 2012).

4.5.2 *Datasets*

In this subsection, we describe the two kinds of datasets for single-label classification of musical instruments. First, we present the *solosDb* dataset of Joder, Essid, and Richard (2009), which consists of solo performances. Secondly, we present the *MedleyDB* dataset of Bittner et al. (2014), which consists of multi-track recordings. Lastly, we describe our chosen method of cross-collection evaluation in which *solosDb* is used for training and validation whereas *MedleyDB* is used for testing.

Datasets of solo music

Joder, Essid, and Richard (2009) have made a dataset of over 500 recordings, spanning across 20 instruments, amounting to about 15 hours of audio. Instruments which are rarely found in commercially available solo music, such as trumpet and clarinet, were complemented with a collection of studio recordings made specifically for this dataset. The authors did not document the split they used, nor the origin of each recording. Yet, because multiple songs from a given artist or album were gathered into this dataset, the classification results are likely to be affected by the choice of split, likewise the dataset of Patil and Elhilali (2015). A conservative workaround is to use the dataset of Joder, Essid, and Richard (2009), thereafter called *solosDb*, only at the evaluation stage, and train the classifier on a different dataset. This method, called cross-collection evaluation, has been repeatedly praised in the context of music information retrieval (Bogdanov et al., 2016; Donnelly and Sheppard, 2015; Livshin and Rodet, 2003).

Multi-track datasets

The problem of designing a *bona fide* dataset of continuous recordings was recently circumvented by the development of multi-track datasets (Bittner et al., 2014; De Man et al., 2014; Fritsch and Plumbley, 2013). In a recording studio, each instrument is assigned to a different stem of microphones, which are subsequently mixed together into a master track by the sound engineer. Releasing these stems for research purposes allows to model the sound of musical instruments independently. The main advantage of this approach is that the scope of available data is not restricted by the repertoire of solo music. In particular, the MedleyDB dataset consists of 122 multi-track songs, covering a broad variety of genres (Bittner et al., 2014). Instrument activations are provided as functions of time, and follow a taxonomy of over 100 elements. A second version of the MedleyDB dataset, with twice as much songs, is forthcoming (Salamon, 2016).

Cross-collection evaluation

In order to evaluate the relative performance of scattering representations, we trained classifiers on a subset of MedleyDB v1.1, and tested them on the solosDb dataset. The subset of musical instruments was chosen so as to maximize the quantity of available data while spanning a relatively wide portion of the organological taxonomy and avoiding morphological confusions (Kolozali et al., 2011; Peeters and Rodet, 2003).

We discarded some prominent instruments in MedleyDB, for various reasons: drums, percussion, and electric bass, because their classification is not challenging enough; male singer, because of falsetto; clean electric guitar, because of audio effects; vocalists, acoustic guitar, and cello, because their morphology was prone to confusion with other classes, namely female singer, electric guitar, and violin. Moreover, we discarded recordings with extended instrumental techniques, since they are extremely rare in MedleyDB.

	training	validation	test
piano	20	8	15
violin	10	4	22
dist. guitar	10	4	11
female singer	7	4	12
B♭ clarinet	5	2	18
flute	3	2	29
B♭ trumpet	4	2	27
tenor sax.	2	1	5
total	61	27	139

Table 4.1: Number of songs per instrument in the proposed dataset of musical instruments. The training set and validation set are derived from MedleyDB. The test set is derived from MedleyDB for distorted electric guitar and female singer, and from solosDb for other instruments.

After performing the automatic detection and segmentation, we personally listened to the dataset entirely. We checked for possible mislabelings, stem bleed, and insufficient sound levels. The problematic situations were reported to the original data providers. The resulting dataset is presented in Table 4.1. It is made publicly available online, along with a version control system [1]. An informal attempt to classify instruments by ear on this dataset confirmed that the human performance on these excerpts is nearly perfect, hence confirming the fairness of the proposed evaluation setting.

1 https://github.com/lostanlen/medleydb-single-instruments

4.5.3 *Methods*

In this subsection, we describe the benchmark of scattering representations which are compared on the same task of musical instrument classification. We study the impact of the choice of convolutional operator (temporal, time-frequency, or spiral), the choice of temporal wavelet (Morlet or Gammatone), and the influence of logarithmic compression.

Pipeline

The dataset presented above is used as a benchmark for the comparison of scattering representations for single-label musical instrument recognition in continuous streams. We investigate three architectures: temporal scattering (Section 3.1), time-frequency scattering (Section 3.2), and spiral scattering (Section 4.2). Moreover, we evaluate the impact of replacing symmetric Morlet wavelets by asymmetric Gammatone wavelets (Subsection 2.2.4) at either or both orders of the wavelet transforms along time. Lastly, we systematically measure the impact of logarithmic compression. Like in the previous experiments on environmental sound classification (Section 3.4), we use a support vector machine (SVM) to perform the classification of scattering features, whose parameters were optimized on a hold-out validation set.

In all experiments, the quality factor of the first-order wavelet filter bank is set to $Q_{\max} = 16$. The center frequencies of auditory wavelets ranges between 83 Hz and 11 kHz, i.e. a range of 7 octaves. Second-order scattering filter banks have a quality factor equal to 1 for all transformed variables.

Finite differences across octaves

Unlike time t and log-frequency γ, the integer-valued octave variable ρ is intrinsically a discrete quantity, which cannot be oversampled. Furthermore, it can only take as many values as octaves in the hearing range, i.e. about $2^3 = 8$. This is much smaller than the typical number of values taken by t and γ, i.e. respectively 2^{16} and 2^7. As a consequence, analytic Morlet wavelets, which were used for visualization of the nonstationary source-filter model (Subsection 4.3.4), might blur out the pitch height of musical instruments, as they have an unbounded support.

For classification purposes, we choose to replace Morlet wavelets by finite differences, which perform additions and subtractions of time-frequency scattering coefficients over a bounded range of three octaves. In doing so, we favor compact support over Heisenberg time-

frequency tradeoff. Within a discrete framework, a family of three finite differences is defined by combining Diracs

$$\psi_{\gamma::\rho}(\rho) = \begin{cases} \frac{1}{2}(\quad \delta(\rho+1) \quad -2\delta(\rho) \quad +\delta(\rho-1)) & \text{if } \gamma::\rho = 2, \\ \frac{1}{\sqrt{2}}(\quad -\delta(\rho+1) \qquad\qquad +\delta(\rho-1)) & \text{if } \gamma::\rho = 1, \\ \frac{1}{\sqrt{3}}(\quad \delta(\rho+1) \quad +\delta(\rho) \quad +\delta(\rho-1)) & \text{if } \gamma::\rho = 0. \end{cases}$$

over the samples -1, 0 and 1. Observe that the filters $\psi_{\gamma::\rho}(\rho)$ are not wavelets in the strict sense, as they do not proceed from each other by dilation. However, they are comparable to wavelets in that they produce an orthonormal basis of $L^2(\{-1,0,1\},\mathbb{R})$ and are differential operators. Indeed, the scale $\gamma::\rho = 2$ corresponds to a moving average with a rectangular window; the scale $\gamma::\rho = 1$, to a discrete first derivative; and the scale $\gamma::\rho = 0$, to a discrete second derivative. Another difference from Morlet wavelets is that the filters $\psi_{\gamma::\rho}(\rho)$, thereafter called finite impulse responses (FIR), are real-valued instead of being analytic. Therefore, the sign variable $\theta::\rho$ may be discarded, as it necessarily takes a null value.

4.5.4 *Results*

In this subsection, we report the accuracies of various methods for the classification of musical instruments in continuous recordings. We find that spiral scattering reaches state-of-the-art results, outperforming time-frequency scattering and deep convolutional networks.

Baseline

Our baseline consists of a typical "bag-of-features" representation, alike the methodologies of Eronen and Klapuri (2000) and Joder, Essid, and Richard (2009). These features consist of the means and standard deviations of spectral shape descriptors, i.e. centroid, bandwidth, skewness, and rolloff; the mean and standard deviation of the zero-crossing rate in the time domain; and the means of mel-frequency cepstral coefficients (MFCC) as well as their first and second derivatives. We train a random forest classifier of 100 decision trees on the resulting feature vector of dimension 70 with balanced class probability.

We obtain an average miss rate of 38.6% across all ten classes. The miss rate is highly variable according to the quantity of data available in each class: although it is excellent for piano (0.3%, 20 tracks) and distorted electric guitar (7.3%, 10 tracks), it is poor for flute (67.5%, 3 tracks) and tenor saxophone (95.6%, 2 tracks).

State of the art excluding scattering

The state of the art in the task of musical instrument classification in continuous recordings is held by deep convolutional networks (ConvNets) (Li, Qian, and Wang, 2015; McFee, Humphrey, and Bello, 2015). Interestingly, many research teams in music information retrieval have converged to employ the same architecture, consisting of two convolutional layers and two densely connected layers (Dieleman and Schrauwen, 2014; Humphrey, Cho, and Bello, 2012; Kereliuk, Sturm, and Larsen, 2015; Li, Qian, and Wang, 2015; McFee, Humphrey, and Bello, 2015; Schlüter and Böck, 2014; Ullrich, Schlüter, and Grill, 2014). However, there is no clear consensus regarding the shape of learned convolutional operators that should be applied to musical audio streams: one-dimensional and two-dimensional receptive fields coexist in the recent literature. The former is comparable to temporal scattering, whereas the latter is comparable to time-frequency scattering.

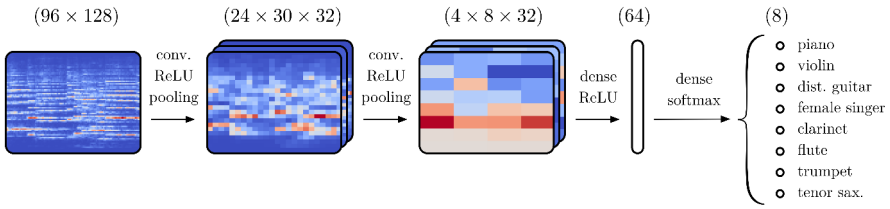


Figure 4.21: A deep convolutional network for musical instrument recognition (Lostanlen and Cella, 2016). Darker red (resp. red) colors denote a larger positive (resp. negative) value.

The average miss rate obtained with a deep convolutional network performing two-dimensional convolutions over the whole time-frequency domain is equal to 30.9% for 32 kernels per layer (93k parameters in total, as shown in Figure 4.21) and 28.3% for 48 kernels per layer (158k parameters in total). Increasing the number of kernels even more causes the accuracy to level out and the variance between trials to increase.

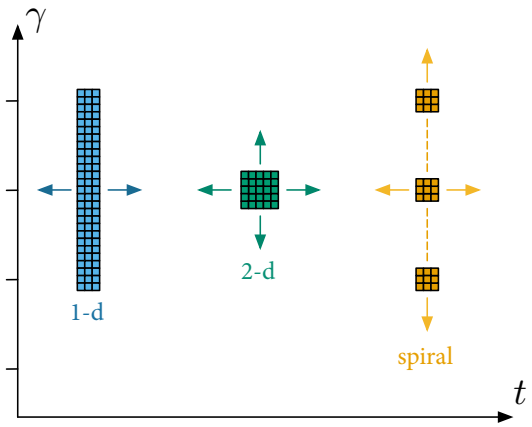


Figure 4.22: Convolutional receptive fields in a neural network. From left to right: one-dimensional convolutions; two-dimensional convolutions; spiral convolutions.

The accuracy of deep convolutional networks for musical instrument classification is improved in our paper (Lostanlen and Cella, 2016) by separating the scalogram into two regions: low frequencies below 2 kHz and high frequencies above 2 kHz, training separate convolutional networks with limited weight sharing (Abdel-Hamid et al., 2014), and hybridizing the outputs of these networks at a deeper stage. Recalling Subsection 4.1.1, the cross-correlations between musical spectra follow a helical structure below 2 kHz and have no rectilinear structure above 2 kHz. Thus, our hybrid network performs temporal convolutions at high frequencies, spiral convolutions at low frequencies, and time-frequency convolutions at all frequencies.

The hybrid network reaches an average miss rate of 26.0% with 32 kernels per layer (147k parameters). This is better than the performance obtained with a conventional architecture of 48 spectrotemporal kernels per layer, with a comparable number of parameters. We refer to our paper for further insight on the construction of spiral convolutional networks (Lostanlen and Cella, 2016).

Performance of scattering representations

The classification accuracies of various scattering representations are reported in Table 4.2.

First of all, like in the environmental sound classification experiments of Section 3.4, the crucial role of logarithmic compression is confirmed. It decreases by approximately one third the average miss rate of all pipelines.

Secondly, time-frequency scattering with Morlet wavelets and logarithmic compression reaches an average miss rate of 22.0%. This figure outperforms the state of the art, held by deep convolutional networks on the Shepard pitch spiral (28.3%). However, replacing time-

frequency scattering by temporal scattering yields an average miss rate of 31.34%, which lags behind the state of the art.

Thirdly, replacing time-frequency scattering by spiral scattering provides a small, albeit consistent boost in classification accuracy, whatever be the chosen wavelet filter banks. The lowest average miss rate achieved in this benchmark is equal to 19.9%, and corresponds to spiral scattering with Morlet wavelets at both orders and logarithmic compression.

Fourthly, replacing Morlet wavelets by Gammatone wavelets at the first order of the scattering transform provides a slight improvement in the case of temporal scattering and time-frequency scattering, but no improvement in the case of spiral scattering. The lowest average miss rate achieved by time-frequency scattering is equal to 20.9%, and corresponds to Gammatone wavelets at the first order, Morlet wavelets at the second order, and logarithmic compression.

Table 4.2: Results in musical instrument classification.

order 1	order 2 (time)	order 2 (log-frequency)	order 2 (octaves)	compression	average miss rate
Morlet	Morlet				38.60
Morlet	Morlet			log	31.98
Gammatone	Gammatone				39.55
Gammatone	Gammatone			log	31.38
Gammatone	Morlet				37.51
Gammatone	Morlet			log	**31.34**
Morlet	Morlet	Morlet			31.11
Morlet	Morlet	Morlet		log	21.97
Gammatone	Gammatone	Morlet			31.11
Gammatone	Gammatone	Morlet		log	20.90
Gammatone	Morlet	Morlet			31.16
Gammatone	Morlet	Morlet		log	**20.89**
Morlet	Morlet	Morlet	FIR		32.93
Morlet	Morlet	Morlet	FIR	log	**19.89**
Gammatone	Gammatone	Morlet	FIR		33.54
Gammatone	Gammatone	Morlet	FIR	log	20.02
Gammatone	Morlet	Morlet	FIR		32.23
Gammatone	Morlet	Morlet	FIR	log	20.39

5

CONCLUSION

In this dissertation, we have presented two new convolutional operators in the time-frequency domain, respectively named time-frequency scattering and spiral scattering, with applications in audio classification. The key idea is to compose wavelet transforms along multiple variables of the scalogram, apply complex modulus, and perform temporal averaging. Both time-frequency scattering and spiral scattering belong to a general framework, called multivariable scattering, which also contains roto-translation scattering in image classification (Oyallon and Mallat, 2015; Sifre and Mallat, 2013).

Audio classification can be formulated as a problem of temporal integration at a certain time scale T. The main challenge addressed by this dissertation is to increase T as much as possible without losing the discriminative patterns in a signal. Whereas T is of the order of 30 ms in speech recognition, it reaches a typical duration of 3 seconds in the case of environmental sound classification (Section 3.4) and musical instrument recognition (Section 4.5), and 30 seconds in the case of acoustic scene classification (Section 3.4).

The recent surge of deep learning systems in audio signal processing (Humphrey, Bello, and Le Cun, 2013) has led to major improvements in classification accuracy with respect to the tradition of engineered features, such as mel-frequency cepstral coefficients (Aucouturier and Pachet, 2004; Eronen and Klapuri, 2000). Yet, the current state of the art in deep learning faces a number of limitations which hinder its adoption by the signal processing community at large. First, the theoretical understanding of deep neural networks in general (Choromanska et al., 2015; Giryes, Sapiro, and Bronstein, 2015), and deep convolutional networks in particular (Mallat, 2016), is still in its infancy. Secondly, few publications have addressed the problem of re-synthesizing sound from deep learning systems (Choi et al., 2015). Thirdly, deep convolutional networks rely on strong assumptions on the data, such as stationarity of statistics and locality of correlations, which are not satisfied by the scalograms of natural sounds (Lostanlen and Cella, 2016).

The multivariable scattering networks introduced in this dissertation can be viewed as deep convolutional networks with complex modulus nonlinearities, in which the convolutional kernels are set equal to multivariable wavelets instead of being learned from annotated data. As such, they provide a model for the shallower layers of a deep convolutional network, with three advantages: mathematical interpretability (Mallat, 2012); the possibility of re-synthesizing sound

by phase retrieval (Mallat, 2016); and high classification accuracy in the small data regime (Bruna and Mallat, 2013b).

In this conclusion, we summarize the content of the dissertation in a way that could hopefully be profitable to both the audio signal processing community and the deep learning community.

5.1 SUMMARY OF FINDINGS

5.1.1 *Multiresolution spectrotemporal analysis*

Besides its biological plausibility (Patil et al., 2012), time-frequency scattering (Section 3.2) overcomes the limitations of temporal scattering, such as the insensitivity to frequency-dependent time shifts (Subsections 3.1.4 and 3.2.3) and the inability to capture coherent time-frequency patterns (Subsections 2.2.3, 3.2.3, and 3.3.3). Because it computes spectrotemporal modulations instead of temporal modulations, it yields a representation of greater dimensionality in feature space, but also of greater regularity through time. As a consequence, one may average time-frequency scattering features over greater durations T than temporal scattering, which leads to a reduced bias in acoustic scene classification (Subsection 3.4.4). Therefore, despite the fact that all multivariable scattering operators presented in this dissertation consist of merely two complex modulus nonlinearities, time-frequency scattering is arguably a deeper model than temporal scattering, because it encompasses more different time scales, from one millisecond to several seconds.

In short, scattering the scalogram into a large number of spectrotemporal scales allows to integrate a broader temporal context than previously existing audio descriptors.

5.1.2 *Audio texture synthesis*

The problem of texture synthesis can be viewed as the characterization of a stationary process from a single realization (Subsections 2.2.3 and 2.4.1). The previous state of the art in texture synthesis is comparable to a temporal scattering transform, supplemented with cross-correlations between neighboring frequencies and high-order statistical moments (McDermott and Simoncelli, 2011). In Subsection 3.3.3, we found that time-frequency scattering, unlike temporal scattering, matches the state of the art in terms of perceptual resemblance to the original. Moreover, because it relies purely on empirical averages and not higher-order moments, time-frequency scattering has the additional benefit of being stable to small deformations. Spiral scattering provides a slight improvement in the reconstruction of polyphonic music with respect to time-frequency scattering (Subsection 4.4.2). However, beyond an averaging duration T of 500 ms, none

of the multivariable scattering representations presented in this dissertation manages to recover intelligible speech, music, or bioacoustic signals.

In short, despite being constrained to a subjective evaluation, the resynthesis methodology adopted in this dissertation allows to examine the discriminative power of an invariant representation on a specific kind of signal, for a specific temporal context T, without classifier or dataset bias.

5.1.3 *Octave equivalence*

Octave equivalence is the fact that two musical notes which are one octave apart in pitch are assigned the same name. It is strongly supported by ethnomusicology (Nettl, 1956), auditory neuroscience (Briley, Breakey, and Krumbholz, 2013), and music psychology (Deutsch, Dooley, and Henthorn, 2008). In order to apply octave equivalence in the time-frequency domain, one may roll up the log-frequency axis into a spiral which makes a full turn at every octave (Subsection 4.1.3). Interestingly, the geometry of the spiral be retrieved from data by manifold learning on the cross-correlations between scalogram features (Subsection 4.1.1). Spiral scattering is derived from this geometry by computing wavelet convolutions across octaves in addition to wavelet convolutions along time and along log-frequency (Section 4.2). In the context of musical instrument classification, spiral scattering outperforms time-frequency scattering (Section 4.5) because it disentangles variations in pitch and variations in timbre (Section 4.3). The same is true of deep convolutional networks (Lostanlen and Cella, 2016).

In short, adapting the shapes of the receptive fields to the geometry of correlations within features improves the generalization power of multivariable scattering representations as well as deep convolutional networks.

5.2 FUTURE PERSPECTIVES

5.2.1 *Large-scale integration*

To perform the numerical experiments presented in this dissertation, we have developed an open source package for multivariable scattering, written in the MATLAB programming language and available at the address: `github.com/lostanlen/scattering.m`. This package is about one order of magnitude faster than the previously available implementation, ScatNet v0.2. It computes a time-frequency scattering transform at ~1x real time on a single core. As a result, datasets like UrbanSound8k or ESC-50 can be analyzed overnight.

Yet, in order to address real-world problems in machine listening, such as automatic generation of social tags in large-scale music corpora (Eck et al., 2007; Oord, Dieleman, and Schrauwen, 2013) or bioacoustic monitoring of wildlife (Salamon et al., 2016), the speed of multivariable scattering transforms should be increased, hopefully by one order of magnitude. To achieve this, we have started developing a new implementation of multivariable scattering, written in the Julia language (Bezanson et al., 2012), available at the address: `github.com/lostanlen/WaveletScattering.jl`.

The Julia implementation performs in-place Fourier transforms and multiplications to avoid allocating memory while relying on multiple dispatch for genericity. Like the MATLAB implementation, it infers the sequence of operations in the scattering network by adopting the "variable as lists" symbolic formalism (Subsection 3.2.1).

Other advantages of migrating to a Julia implementation is that it does not require a license, and can quite easily be linked to a library of low-level routines for deep learning. Furthermore, Julia programs can be ported to the ARM processors of low-cost acoustic monitoring devices (Mydlarz, Salamon, and Bello, 2017).

5.2.2 *Deep learning meets multiresolution analysis*

In this dissertation, we have strived for simplicity by performing all classification experiments with a support vector machine (SVM), a powerful but shallow classifier. In order to improve classification accuracy, scattering features could be fed to a deep neural network instead.

In the context of speech recognition, this has been experimented by Peddinti et al. (2014) and Sainath et al. (2014) and Zeghidour et al. (2016), in supervised and unsupervised settings respectively. Yet, these publications do not take into account the locality of correlations between scattering coefficients in the time-frequency domain. Indeed, they rely on densely connected layers instead of convolutional layers.

In Subsection 3.4.4, we have seen that time-frequency scattering associated with a support vector machine is outperformed by convolutional networks trained on scalograms on a task of acoustic scene classification ($T = 30\,\mathrm{s}$). In this context, future work could address the stacking of a deep convolutional network on top of a time-frequency scattering network.

On the other hand, some recent publications have shown the interest of adopting multiscale approaches to deep learning (Dieleman and Schrauwen, 2013), with successful applications in classification (Hamel, Bengio, and Eck, 2012) as well as re-synthesis (Oord et al., 2016). Therefore, it seems that bridging the gap between scalogram-based deep learning and scattering representations is an area ripe for future research.

5.2.3 *The revolution will not be supervised*

One may argue that deep learning is becoming the central topic in pattern recognition of our decade, in the same way that wavelets and sparse representations respectively emerged as central topics in the 1990s and 2000s. The first and most prominent examples of the success of deep learning in audio classification are English and Mandarin Chinese speech recognition, in which state-of-the-art methods learn features from the raw waveform, thus outperforming scalogram-based features (Sainath et al., 2015). This is because industrial research laboratories have massively invested in these languages, providing thousands of hours of annotated data (Amodei et al., 2016).

Conversely, current datasets of musical instrument sounds are limited to a few hours only. This is due to the cost of multi-track recording and manual annotation. As a result, the classification accuracy of deep convolutional networks trained on the audio waveform is on par with feature engineering, but not substantially better (Li, Qian, and Wang, 2015). Although the use of data augmentation often provides a small boost in accuracy (McFee, Humphrey, and Bello, 2015; Salamon and Bello, 2016), it does not address the fundamental lack of diverse data which prevents the massive adoption of deep learning in musical instrument classification. To mitigate this problem, we are currently taking part in the foundation of a community for open and sustainable music information research (COSMIR), following the guidelines of McFee, Humphrey, and Urbano (2016).

Other problems in music information retrieval, such as automatic tagging, are less well-defined from a taxonomical point of view but benefit from larger datasets. Yet, interestingly, Choi, Fazekas, and Sandler (2016) have shown that switching from the MagnaTagaTune dataset (25k songs) to the Million Song Dataset (Bertin-Mahieux et al., 2011) only improved marginally the classification accuracy of deep convolutional networks. Therefore, it seems that the limited amount of training data is not the only culprit of the limited success of deep learning. Furthermore, as Malcolm Slaney (Google Research) wittingly said in a personal communication, "there are only thirty million songs out there", i.e. on digital music streaming platforms: this is one order of magnitude below the number of pictures uploaded on Facebook during the last decade only. In this context, unsupervised learning of auditory representations from unlabeled videos appears as a promising direction of research (Aytar, Vondrick, and Torralba, 2016).

Is the renewed interest in deep learning the premise of a revolution in artificial intelligence? Despite the fact that scientific revolutions remain invisible as long as they do not cause a thorough update of the classical textbooks in the field (Kuhn, 1962, chapter 9), some influential researchers have firmly answered in the affirmative already (Le Cun, Bengio, and Hinton, 2015; Schmidhuber, 2015). Yet, for the

reasons discussed above, it is unclear when deep learning will match human performance in general-purpose audio classification, if ever. Instead, there are grounds to believe that the convolutional operators presented in this dissertation, which do not need any supervision, will remain relevant in the next years.

It is too early to estimate the impact of the so-called "deep learning revolution". But one thing is certain: as far as acoustic scene analysis is concerned, one may pastiche the famous words of singer-songwriter Gill Scott-Heron, and claim that the revolution will not be supervised.

A

DESIGN OF GAMMATONE WAVELETS

In Subsection 2.2.4, we have presented the Gammatone wavelet $\psi(t)$, a complex-valued, band-pass filter of null mean, center frequency ξ, and bandwidth σ. For some integer $N > 1$, the analytic expression of the Gammatone wavelet is

$$\psi(t) = \left(2\pi(\mathrm{i} - \sigma)t^{N-1} + (N-1)t^{N-2}\right) \exp(-2\pi\sigma t) \exp(2\pi\mathrm{i}\xi t).$$

In this appendix, we provide a rationale for choosing the topmost center frequency ξ of a Gammatone wavelet filter bank in a discrete-time setting. Then, we relate the bandwidth parameter σ to the choice of a quality factor Q.

Center frequency of the mother wavelet

In order to preserve energy and allow for perfect reconstruction, the Gammatone wavelet filter bank must satisfy the inequalities

$$1 - \varepsilon \leq |\hat{\phi}(\omega)| + \sum_{\gamma} |\hat{\psi}(2^{\gamma}\omega)| + |\hat{\psi}(-2^{\gamma}\omega)| \leq 1$$

for all frequencies ω, where ε is a small margin (Andén and Mallat, 2014). Satisfying the equation above near the Nyquist frequency $\omega = \pi$ can be achieved by placing the log-frequency $\log_2 \xi$ of the first (topmost) wavelet in between the log-frequency $\log_2(\xi \times 2^{-1/Q})$ of the second wavelet and the log-frequency $\log_2(2\pi - \xi)$ of the mirror of the first wavelet. We obtain the equation

$$\log_2 \xi - \log_2(\xi \times 2^{-1/Q}) = \log_2(2\pi - \xi) - \log_2 \xi,$$

of which we deduce the identity

$$\xi = \frac{2\pi}{1 + 2^{1/Q}}.$$

For $Q = 1$, this yields a center frequency of $\xi = \frac{2\pi}{3}$. For greater values of Q, the center frequency ξ tends towards π.

Bandwidth parameter

The quality factor Q of the Gammatone wavelet is defined as the ratio between the center frequency ξ of the wavelet $\hat{\psi}(\omega)$ and its bandwidth B in the Fourier domain. This bandwidth is given by the difference between the two solutions ω of the following equation:

$$\frac{|\hat{\boldsymbol{\psi}}(\omega)|}{|\hat{\boldsymbol{\psi}}(\xi)|} = \frac{\omega}{\xi} \times \left(1 + \frac{(\omega - \xi)^2}{\sigma^2}\right)^{-N/2} = r,$$

where the magnitude cutoff r is most often set to $\sqrt{\frac{1}{2}}$. Let $\Delta\omega = \omega - \xi$. Raising the above equation to the power $N/2$ yields the following:

$$\left(1 + \frac{\Delta\omega}{\omega_c}\right)^{N/2} = r \times \left(1 + \frac{\Delta\omega^2}{\alpha^2}\right).$$

Since $\Delta\omega \ll \xi$, we may approximate the left-hand side with a first-order Taylor expansion. This leads to a quadratic equation of the variable $\Delta\omega$:

$$\frac{r^{2/N}}{\sigma^2} \times \Delta\omega^2 - \frac{2}{N\xi} \times \Delta\omega + \left(r^{2/N} - 1\right) = 0.$$

The discriminant of the above equation is:

$$D = 4 \times \left(\frac{1}{N^2\xi^2} + \frac{r^{2/N}\left(1 - r^{2/N}\right)}{\sigma^2}\right),$$

which is a positive number as long as $r < 1$. The bandwidth B of $\hat{\boldsymbol{\psi}}$ is given by the difference between the two solutions of the quadratic equation, that is:

$$B = \frac{2\sigma^2}{r^{2/N}} \times \sqrt{\frac{1}{N^2\xi^2} + \frac{r^{2/N}\left(1 - r^{2/N}\right)}{\sigma^2}}.$$

Now, let us express the parameter α as a function of some required bandwidth B at some cutoff threshold r. After having raised the above to its square and rearranged the terms, we obtain another quadratic equation, yet of the variable α^2:

$$\frac{4}{r^{4/N}N^2\xi^2}\sigma^4 + \frac{4 \times \left(1 - r^{2/N}\right)}{r^{2/N}}\sigma^2 - B^2 = 0$$

We multiply the equation by $\dfrac{r^{2/N}}{4 \times \left(1 - r^{2/N}\right)} \neq 0$:

$$\frac{1}{r^{2/N}\left(1 - r^{2/N}\right)N^2\xi^2}\sigma^4 + \sigma^2 - \frac{r^{2/N}B^2}{4 \times \left(1 - r^{2/N}\right)} = 0$$

This leads to defining σ^2 as the unique positive root of the above polynomial:

$$\sigma^2 = \frac{r^{2/N}\left(1 - r^{2/N}\right)N^2\xi^2}{2} \times \left(\sqrt{1 + \frac{B^2}{\left(1 - r^{2/N}\right)^2 N^2\xi^2}} - 1\right).$$

If the filter bank has to be approximately orthogonal, we typically set B to $B = \left(1 - 2^{-1/Q}\right) \times \xi$.

We conclude with the following closed form for α:

$$\alpha = K_N \times \sqrt{\frac{\sqrt{1 + h_N(Q)^2} - 1}{2}} \times \xi,$$

where

$$K_N = r^{1/N} N \sqrt{1 - r^{2/N}} \text{ and } h_N(Q) = \frac{1 - 2^{-1/Q}}{N \times (1 - r^{2/N})}.$$

B

PROOF OF EQUATION 4.6

Theorem 1. *Let $h \in \mathbf{L}^2(\mathbb{R})$ of finite support T_h and $\eta \in \mathcal{C}^2(\mathbb{R})$ such that $\dot{\eta}(t) > 0$. The wavelet transform of $h_\eta(t) = \dot{\eta}(t) \times (h \circ \eta)(t)$ with the wavelet $\psi_{2^\gamma}(t) = g(2^\gamma t) \times \exp(\mathrm{i}2\pi 2^\gamma t)$ is equal to*

$$(h_\eta * \psi_{2^\gamma})(t) = \psi_{2^\gamma}\left(\frac{\eta(t)}{\dot{\eta}(t)}\right) \times \widehat{h}\left(\frac{2^\gamma}{\dot{\eta}(t)}\right) + \varepsilon(t,\gamma).$$

The corrective term satisfies

$$\begin{aligned}\frac{\varepsilon(t,\gamma)}{\|h\|_\infty \times T_h} &\leq g\left(2^\gamma \frac{\eta(t)}{\dot{\eta}(t)}\right) \times T_h^2 \times \left\|\frac{\ddot{\eta}}{\dot{\eta}^3}\right\|_{t,T_h} \\ &+ 2^\gamma \dot{g}\left(2^\gamma \frac{\eta(t)}{\dot{\eta}(t)}\right) \times \left\|\frac{1}{\dot{\eta}(t)}\right\|_{t,T_h} \\ &+ 2^{2\gamma} \times \left\|\frac{1}{\dot{\eta}}\right\|_{t,T_h}^2 \times \|\ddot{g}\|_\infty\end{aligned}$$

where $\|\cdot\|_{t,T_h}$ is a shorthand for the supremum norm over the support of h:

$$\|x\|_{t,T_h} = \sup_{|u-t|<T_h} |x(u)|.$$

Proof. The convolution between the deformed filter and the wavelet is defined as

$$(h_\eta * \psi_{2^\gamma})(t) = \int_{-\infty}^{+\infty} \dot{\eta}(u) \times (h \circ \eta)(u) \times \psi_{2^\gamma}(t-u)\,\mathrm{d}u.$$

The change of variables $v = \eta(u)$ leads to

$$(h_\eta * \psi_{2^\gamma})(t) = \int_{-\infty}^{+\infty} h(v) \times \psi_\gamma\left(t - \eta^{(-1)}(v)\right)\,\mathrm{d}v,$$

where $\eta^{(-1)}$ is the functional inverse of η. We compute the second-order Taylor expansion of $\eta^{(-1)}$ around $\eta(t)$:

$$\eta^{(-1)}(v) = t + \frac{(v-\eta(t))}{\dot{\eta}(t)} + \frac{(v-\eta(t))^2}{2}\varepsilon_{\eta^{(-1)}}(v)$$

with

$$\left|\varepsilon_{\eta^{(-1)}}(v)\right| \leq \sup_{\left[t,\eta^{(-1)}(v)\right]} \left|\frac{\ddot{\eta}}{\dot{\eta}^3}\right|.$$

The wavelet ψ_{2^γ} is decomposed into analytic amplitude and phase

$$\psi_{2^\gamma}\left(t - \eta^{(-1)}(v)\right) = g\left(2^\gamma(t-\eta^{(-1)}(v))\right)\exp\left(2\pi\mathrm{i}2^\gamma(t-\eta^{(-1)}(v))\right).$$

Let us define the auxiliary function

$$\boldsymbol{\rho}(v) = -\frac{v}{\dot{\boldsymbol{\eta}}(t)} - \frac{(v-\boldsymbol{\eta}(t))^2}{2}\boldsymbol{\varepsilon}_{\boldsymbol{\eta}^{(-1)}}(v)\,. \tag{B.1}$$

We have

$$g\left(2^\gamma(t-\boldsymbol{\eta}^{(-1)}(v))\right) = g\left(2^\gamma\left(\frac{\boldsymbol{\eta}(t)}{\dot{\boldsymbol{\eta}}(t)} + \boldsymbol{\rho}(v)\right)\right).$$

We expand the envelope g again with a second-order Taylor expansion around $\boldsymbol{\eta}(t)$ according to the variable $\boldsymbol{\rho}(v)$

$$\begin{aligned} g\left(2^\gamma(t-\boldsymbol{\eta}^{(-1)}(v))\right) &= g\left(2^\gamma\frac{\boldsymbol{\eta}(t)}{\dot{\boldsymbol{\eta}}(t)}\right) \\ &+ 2^\gamma\dot{g}\left(2^\gamma\frac{\boldsymbol{\eta}(t)}{\dot{\boldsymbol{\eta}}(t)}\right)\times\boldsymbol{\rho}(v) \\ &+ \frac{2^{2\gamma}\boldsymbol{\rho}(v)^2}{2}\boldsymbol{\varepsilon}_g\left(\boldsymbol{\rho}(v)\right) \end{aligned}$$

and make the crude majorization

$$\left|\boldsymbol{\varepsilon}_g\left(\boldsymbol{\rho}(v)\right)\right| \leq \sup_{\mathbb{R}}|\ddot{\boldsymbol{g}}|.$$

We obtain

$$\begin{aligned} &\left(\boldsymbol{h}_{\boldsymbol{\eta}} * \boldsymbol{\psi}_{2^\gamma}\right)(t)\times\exp\left(-2\pi\mathrm{i}2^\gamma\frac{\boldsymbol{\eta}(t)}{\dot{\boldsymbol{\eta}}(t)}\right) \\ =\quad & g\left(2^\gamma\frac{\boldsymbol{\eta}(t)}{\dot{\boldsymbol{\eta}}(t)}\right)\int_{-\infty}^{+\infty}\boldsymbol{h}(v)\exp\left(-2\pi\mathrm{i}\frac{2^\gamma}{\dot{\boldsymbol{\eta}}(t)}v\right) \\ &\exp\left(-2\pi\mathrm{i}2^\gamma\frac{(v-\boldsymbol{\eta}(t))^2}{2}\boldsymbol{\varepsilon}_{\boldsymbol{\eta}^{(-1)}}(v)\right)\,\mathrm{d}v \qquad \text{(I)} \\ +\quad & 2^\gamma\dot{g}\left(2^\gamma\frac{\boldsymbol{\eta}(t)}{\dot{\boldsymbol{\eta}}(t)}\right)\int_{-\infty}^{+\infty}\boldsymbol{\rho}(v)\boldsymbol{h}(v)\exp\left(-2\pi\mathrm{i}\frac{2^\gamma}{\dot{\boldsymbol{\eta}}(t)}v\right) \\ &\exp\left(-2\pi\mathrm{i}2^\gamma\frac{(v-\boldsymbol{\eta}(t))^2}{2}\boldsymbol{\varepsilon}_{\boldsymbol{\eta}^{(-1)}}(v)\right)\,\mathrm{d}v \qquad \text{(II)} \\ +\quad & \frac{2^{2\gamma}}{2}\int_{-\infty}^{+\infty}\boldsymbol{\rho}(v)^2\boldsymbol{\varepsilon}_g(\boldsymbol{\rho}(v))\boldsymbol{h}(v) \\ &\exp\left(-2\pi\mathrm{i}2^\gamma\boldsymbol{\rho}(v)\right)\,\mathrm{d}v. \qquad \text{(III)} \end{aligned}$$

A first-order Taylor expansion of the complex exponential gives

$$\exp\left(-2\pi\mathrm{i}2^\gamma\frac{v^2}{2}\boldsymbol{\varepsilon}_{\boldsymbol{\eta}^{(-1)}}(v)\right) \approx 1. \tag{B.2}$$

We recognize the Fourier transform of $\boldsymbol{h}(v)$ evaluated at the negative frequency $-2^\gamma/\dot{\boldsymbol{\eta}}(t)$:

$$\int_{-\infty}^{+\infty}\boldsymbol{h}(v)\exp\left(2\pi\mathrm{i}\frac{2^\gamma}{\dot{\boldsymbol{\eta}}(t)}v\right)\,\mathrm{d}v = \widehat{\boldsymbol{h}}\left(-\frac{2^\gamma}{\dot{\boldsymbol{\eta}}(t)}\right).$$

Since $h(v)$ is real, $\hat{h}(\omega)$ is symmetric, so we may drop the sign of $-2^\gamma/\dot{\eta}(t)$. The first term in the expansion rewrites as

$$\text{(I)} \approx g\left(2^\gamma \frac{\eta(t)}{\dot{\eta}(t)}\right) \times \hat{h}\left(\frac{2^\gamma}{\dot{\eta}(t)}\right).$$

Inserting Equation B.2 in the expression of Equation yields

$$\left|\left(h_\eta * \psi_{2^\gamma}\right)(t) - \psi_{2^\gamma}\left(\frac{\eta(t)}{\dot{\eta}(t)}\right)\hat{h}\left(\frac{2^\gamma}{\dot{\eta}(t)}\right)\right| \leq \varepsilon_{\text{(I)}} + \varepsilon_{\text{(II)}} + \varepsilon_{\text{(III)}}$$

where

$$\varepsilon_{\text{(I)}} = g\left(2^\gamma \frac{\eta(t)}{\dot{\eta}(t)}\right)\left|\int_{-\infty}^{+\infty} h(v) v^2 \varepsilon_{\eta^{(-1)}}(v)\,\mathrm{d}v\right|,$$

$$\varepsilon_{\text{(II)}} = 2^\gamma \dot{g}\left(2^\gamma \frac{\eta(t)}{\dot{\eta}(t)}\right)\left|\int_{-\infty}^{+\infty} \rho(v) h(v)\,\mathrm{d}v\right|,$$

and

$$\varepsilon_{\text{(III)}} = \frac{2^{2\gamma}}{2}\left|\int_{-\infty}^{+\infty} \rho(v)^2 \varepsilon_g(\rho(v)) h(v)\,\mathrm{d}v\right|$$

Let T_h be the time support of the filter $h(t)$. We dominate all three corrective terms with a simple "base times height" inequality. The term v^2 is dominated by T_h^2, $h(v)$ is dominated by $\|h\|_\infty$ and $\varepsilon_{\eta^{(-1)}}(v)$ is dominated by $\left\|\frac{\ddot{\eta}}{\dot{\eta}^3}\right\|_{t,T_h}$:

$$\varepsilon_{\text{(I)}} \leq g\left(2^\gamma \frac{\eta(t)}{\dot{\eta}(t)}\right) \times T_h^2 \times \|h\|_\infty \times \left\|\frac{\ddot{\eta}}{\dot{\eta}^3}\right\|_{t,T_h} \times T_h.$$

We have, by first order expansion of $\eta^{(-1)}(v)$ around t:

$$|\rho(v)| \leq \left\|t - \eta^{(-1)}(v)\right\|_\infty \leq \left\|\frac{1}{\dot{\eta}}\right\|_{t,T_h}.$$

Plugging this inequality into the definition of $\varepsilon_{\text{(II)}}$ yields

$$\varepsilon_{\text{(II)}} \leq 2^\gamma \dot{g}\left(2^\gamma \frac{\eta(t)}{\dot{\eta}(t)}\right) \times \left\|\frac{1}{\dot{\eta}(t)}\right\|_{t,T_h} \times \|h\|_\infty \times T_h.$$

We dominate $\varepsilon_{\text{(III)}}$ in the same way:

$$\varepsilon_{\text{(III)}} \leq 2^{2\gamma} \times \left\|\frac{1}{\dot{\eta}}\right\|_{t,T_h}^2 \times \|\ddot{g}\|_\infty \times \|h\|_\infty \times T_h.$$

Let $\varepsilon = \varepsilon_{\text{(I)}} + \varepsilon_{\text{(II)}} + \varepsilon_{\text{(III)}}$ be the full corrective term. We conclude with the desired upper bound. □

BIBLIOGRAPHY

Abdel-Hamid, O et al. (2014). "Convolutional neural networks for speech recognition." In: *IEEE Transactions on Audio, Signal, and Language Processing* 22.10, pp. 1533–1545 (cit. on p. 142).

Aertsen, AMHJ and PIM Johannesma (1981). "The spectro-temporal receptive field." In: *Biological Cybernetics* 42.2, pp. 133–143 (cit. on p. 65).

Agcaer, S et al. (2015). "Optimization of amplitude modulation features for low-resource acoustic scene classification." In: *Proceedings of the European Signal Processing Conference (EUSIPCO)* (cit. on p. 90).

Alam, MJ et al. (2013). "Amplitude modulation features for emotion recognition from speech." In: *Proceedings of INTERSPEECH* (cit. on p. 49).

Amodei, D et al. (2016). "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin." In: (cit. on p. 149).

Andén, J, V Lostanlen, and S Mallat (2015). "Joint time-frequency scattering for audio classification." In: *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (cit. on p. 49).

Andén, J and S Mallat (2012). "Scattering representation of modulated sounds." In: *Proceedings of the International Conference on Digical Audio Effects (DAFx)* (cit. on pp. 57, 124).

— (2014). "Deep scattering spectrum." In: *IEEE Transactions on Signal Processing* 62.16, pp. 4114–4128 (cit. on pp. 6, 36, 49, 50, 52, 54, 57, 58, 67, 81, 87, 151).

Atlas, L and SA Shamma (2003). "Joint acoustic and modulation frequency." In: *EURASIP Journal on Advances in Signal Processing* 2003.7, pp. 1–8 (cit. on pp. 3, 53).

Aucouturier, J and F Pachet (2003). "Representing musical genre: a state of the art." In: *Journal of New Music Research* 32.1, pp. 83–93 (cit. on p. 53).

— (2004). "Improving timbre similarity: how high's the sky?" In: *Journal of Negative Results in Speech and Audio Sciences* 1.1 (cit. on pp. 5, 38, 145).

Auger, F et al. (2013). "Time-frequency reassignment and synchrosqueezing: an overview." In: *IEEE Signal Processing Magazine* 30.6, pp. 32–41 (cit. on p. 24).

Aytar, Y, C Vondrick, and A Torralba (2016). "Learning Sound Representations from Unlabeled Video." In: *Advances in Neural Information Processing Systems (NIPS)* (cit. on p. 149).

Bacry, E, JF Muzy, and A Arnéodo (1993). "Singularity spectrum of fractal signals from wavelet analysis: exact results." In: *Journal of Statistical Physics* 70.3-4, pp. 635–674 (cit. on p. 102).

Bae, SH, I Choi, and NS Kim (2016). "Acoustic scene classification using parallel combination of LSTM and CNN." In: *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)* (cit. on p. 81).

Balazs, P et al. (2011). "Theory, implementation and applications of nonstationary Gabor frames." In: *Journal of Computational and Applied Mathematics* 236.6, pp. 1481–1496 (cit. on pp. 12, 18).

Baraniuk, RG and DL Jones (1992). "New dimensions in wavelet analysis." In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 5 (cit. on p. 64).

Barchiesi, D et al. (2015). "Acoustic scene classification: classifying environments from the sounds they produce." In: *IEEE Signal Processing Magazine* 32.3, pp. 16–34 (cit. on p. 80).

Bartsch, MA and GH Wakefield (2005). "Audio thumbnailing of popular music using chroma-based representations." In: *IEEE Transactions on Multimedia* 7.1, pp. 96–104 (cit. on p. 106).

Baugé, C et al. (2013). "Representing environmental sounds using the separable scattering transform." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 81).

Bello, JP et al. (2005). "A tutorial on onset detection in music signals." In: *IEEE Transactions on Speech and Audio Processing* 13.5, pp. 1035–1047 (cit. on p. 21).

Benetos, E, M Kotti, and C Kotropoulos (2006). "Musical instrument classification using non-negative matrix factorization algorithms." In: *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCS)* (cit. on p. 136).

Bengio, Y (2013). "Deep learning of representations: looking forward." In: *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)* (cit. on p. 3).

Bertin-Mahieux, T and DPW Ellis (2011). "Large-scale cover song recognition using hashed chroma landmarks." In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (cit. on p. 107).

Bertin-Mahieux, T et al. (2011). "The Million Song Dataset." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on p. 149).

Beylkin, G, R Coifman, and V Rokhlin (1991). "Fast wavelet transforms and numerical algorithms I." In: *Communications on Pure and Applied Mathematics* 44.2, pp. 141–183 (cit. on p. 16).

Bezanson, J et al. (2012). "Julia: a fast dynamic language for technical computing." In: *arXiv preprint 1209.5145* (cit. on p. 148).

Bhalke, DG, CBR Rao, and DS Bormane (2016). "Automatic musical instrument classification using fractional Fourier transform-based-MFCC features and counter-propagation neural network." In: *Journal of Intelligent Information Systems* 46.3, pp. 425–446 (cit. on p. 136).

Bittner, RM et al. (2014). "MedleyDB: a multitrack dataset for annotation-intensive MIR research." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on pp. 8, 137).

Blauert, Jens (2004). *Spatial Hearing: The psychophysics of human sound localization (revised edition)*. Cambridge, MA, USA: The MIT Press (cit. on p. 85).

Bogdanov, D et al. (2016). "Cross-collection evaluation for music classification tasks." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on p. 137).

Bogert, Bruce P, Michael JR Healy, and John W Tukey (1963). "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking." In: *Proceedings of the Symposium on Time Series Analysis*. Vol. 15 (cit. on p. 36).

Braus, Ira (1995). "Retracing One's Steps: An Overview of Pitch Circularity and Shepard Tones in European Music, 1550–1990." In: *Music Perception: An Interdisciplinary Journal* 12.3, pp. 323–351 (cit. on p. 103).

Briley, Paul M, Charlotte Breakey, and Katrin Krumbholz (2013). "Evidence for pitch chroma mapping in human auditory cortex." In: *Cerebral Cortex* 23.11, pp. 2601–2610 (cit. on pp. 95, 147).

Brown, Judith C (1999). "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features." In: *Journal of the Acoustical Society of America* 105.3, pp. 1933–1941 (cit. on pp. 38, 135).

Brown, Judith C and Patrick JO Miller (2007). "Automatic classification of killer whale vocalizations using dynamic time warping." In: *Journal of the Acoustical Society of America* 122.2, pp. 1201–1207 (cit. on p. 34).

Bruna, J and S Mallat (2013a). "Audio texture synthesis with scattering moments." In: *arXiv preprint 1311.0407* (cit. on pp. 70, 72).

— (2013b). "Invariant scattering convolution networks." In: *IEEE Transactions on Pattern Analysis and Machine intelligence* 35.8, pp. 1872–1886 (cit. on p. 146).

Bruna, J, P Sprechmann, and Y Le Cun (2015). "Source separation with scattering non-negative matrix factorization." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 55).

Bultan, Aykut (1999). "A four-parameter atomic decomposition of chirplets." In: *IEEE Transactions on Signal Processing* 47.3, pp. 731–745 (cit. on p. 64).

Burns, Edward M. (1999). "Intervals, Scales, and Tuning." In: *The Psychology of Music*. Second Edition. Cognition and Perception. Academic Press, pp. 215–264 (cit. on p. 95).

Cano, P et al. (2005). "A Review of Audio Fingerprinting." In: *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology* 41.3, pp. 271–284 (cit. on p. 53).

Casati, R and J Dokic (1994). *La philosophie du son*. Jacqueline Chambon (cit. on p. 43).

Casey, Michael (2001). "MPEG-7 sound-recognition tools." In: *IEEE Transactions on circuits and Systems for video Technology* 11.6, pp. 737–747 (cit. on p. 136).

Castellengo, M and D Dubois (2007). "Timbre ou timbres? Propriété du signal, de l'instrument, ou construction cognitive?" In: *Cahiers de la Société québécoise de recherche en musique* 9, pp. 25–38 (cit. on p. 44).

Chachada, S and CCJ Kuo (2014). "Environmental sound recognition: a survey." In: *APSIPA Transactions on Signal and Information Processing* 3 (cit. on p. 80).

Chang, CC and CJ Lin (2011). "LIBSVM: a library for support vector machines." In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3, p. 27 (cit. on p. 88).

Cheveigné, A de (2005). "Pitch perception." In: *Oxford Handbook of Auditory Science: Hearing*. Oxford University Press. Chap. 4, pp. 71–104 (cit. on p. 39).

Chi, T, P Ru, and SA Shamma (2005). "Multiresolution spectrotemporal analysis of complex sounds." In: *Journal of the Acoustical Society of America* 118.2, pp. 887–906 (cit. on pp. 4, 6, 49, 59, 66).

Cho, T and JP Bello (2014). "On the relative importance of individual components of chord recognition systems." In: *IEEE Transactions on Audio, Speech, and Language Processing* 22.2, pp. 477–492 (cit. on p. 107).

Choi, K, G Fazekas, and M Sandler (2016). "Automatic tagging using deep convolutional neural networks." In: *Proceedings of the International Society of Music Information Retrieval (ISMIR)* (cit. on pp. 5, 149).

Choi, K et al. (2015). "Auralisation of deep convolutional neural networks: listening to learned features." In: *Proceedings of the International Society of Music Information Retrieval (ISMIR)* (cit. on p. 145).

Choromanska, A et al. (2015). "The Loss Surfaces of Multilayer Networks." In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)* (cit. on p. 145).

Chu, S, S Narayanan, and CCJ Kuo (2009). "Environmental sound recognition with time-frequency audio features." In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.6, pp. 1142–1158 (cit. on p. 80).

Cohen, L (1995). *Time-frequency analysis*. Prentice Hall (cit. on p. 29).

Conan, S et al. (2013). "Navigating in a space of synthesized interaction-sounds: rubbing, scratching and rolling sounds." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 27).

Cooley, JW and John W Tukey (1965). "An algorithm for the machine calculation of complex Fourier series." In: *Mathematics of computation* 19.90, pp. 297–301 (cit. on p. 16).

Cortes, C and V Vapnik (1995). "Support-vector networks." In: *Machine Learning* 20.3, pp. 273–297 (cit. on p. 88).

Damoulas, T et al. (2010). "Bayesian classification of flight calls with a novel dynamic time warping kernel." In: *International Conference on Machine Learning and Applications (ICMLA)*. IEEE (cit. on p. 38).

Dau, T, B Kollmeier, and A Kohlrausch (1997). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers." In: *Journal of the Acoustical Society of America* 102.5, pp. 2892–2905 (cit. on p. 53).

Daubechies, I (1996). "Where do wavelets come from? A personal point of view." In: *Proceedings of the IEEE* 84.4, pp. 510–513 (cit. on p. 14).

Davis, S and P Mermelstein (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4, pp. 357–366 (cit. on p. 38).

De Boer, E and P Kuyper (1968). "Triggered correlation." In: *IEEE Transactions on Biomedical Engineering* 3, pp. 169–179 (cit. on p. 65).

De Cheveigné, A and H Kawahara (2002). "YIN, a fundamental frequency estimator for speech and music." In: *Journal of the Acoustical Society of America* 111.4, pp. 1917–1930 (cit. on p. 37).

De Man, Brecht et al. (2014). "The Open Multitrack Testbed." In: *Proceedings of the Audio Engineering Society Convention* (cit. on p. 137).

Delprat, N et al. (1992). "Asymptotic Wavelet and Gabor Analysis: Extraction of Instantaneous Frequencies." In: *IEEE Transactions on Information Theory* 38, pp. 644–664 (cit. on pp. 19, 26, 68, 116, 124, 125).

Depireux, DA et al. (2001). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex." In: *Journal of Neurophysiology* 85.3, pp. 1220–1234 (cit. on p. 66).

Deutsch, D (1994). "The tritone paradox: Some further geographical correlates." In: *Music Perception: An Interdisciplinary Journal* 12.1, pp. 125–136 (cit. on p. 103).

Deutsch, D, K Dooley, and T Henthorn (2008). "Pitch circularity from tones comprising full harmonic series." In: *Journal of the Acoustical Society of America* 124.1, pp. 589–597 (cit. on pp. 99, 147).

Dieleman, S and B Schrauwen (2014). "End-to-end learning for music audio." In: *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)* (cit. on p. 141).

Dieleman, Sander and Benjamin Schrauwen (2013). "Multiscale approaches to music audio feature learning." In: *14th International Society for Music Information Retrieval Conference (ISMIR)* (cit. on p. 148).

Donnelly, PJ and JW Sheppard (2015). "Cross-Dataset Validation of Feature Sets in Musical Instrument Classification." In: *Proceedings of the IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE (cit. on p. 137).

Eck, D et al. (2007). "Automatic Generation of Social Tags for Music Recommendation." In: *Advances in Neural Information Processing Systems (NIPS)* (cit. on pp. 21, 148).

Eerola, T and R Ferrer (2008). "Instrument library (MUMS) revised." In: *Music Perception: An Interdisciplinary Journal* 25.3, pp. 253–255 (cit. on p. 135).

Eggermont, JJ (1993). "Wiener and Volterra analyses applied to the auditory system." In: *Hearing research* 66.2, pp. 177–201 (cit. on p. 65).

Eghbal-Zadeh, H et al. (2016). "A hybrid approach using binaural i-vector and deep convolutional neural networks." In: *Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)* (cit. on pp. 80, 91).

Ellis, DPW (2007). "Beat tracking by dynamic programming." In: *Journal of New Music Research* 36.1, pp. 51–60 (cit. on pp. 21, 106).

Ellis, DPW and GE Poliner (2007). "Identifying cover songs with chroma features and dynamic programming beat tracking." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 4 (cit. on p. 107).

Ellis, DPW, X Zeng, and JH McDermott (2011). "Classifying soundtracks with audio texture features." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE (cit. on p. 81).

Emiya, V, R Badeau, and B David (2010). "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle." In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.6, pp. 1643–1654 (cit. on p. 119).

Eronen, A and A Klapuri (2000). "Musical instrument recognition using cepstral coefficients and temporal features." In: *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (cit. on pp. 38, 135, 140, 145).

Esqueda, F and V Välimäki (2015). "Barberpole phasing and flanging illusions." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 111).

Essid, S et al. (2005). "On the usefulness of differentiated transient / steady-state processing in machine recognition of musical instruments." In: *Proceedings of the Convention of the Audio Engineering Society (AES)* (cit. on p. 115).

Fastl, H and E Zwicker (2007). *Psychoacoustics: Facts and models*. Springer (cit. on p. 28).

Faure, A, S McAdams, and V Nosulenko (1996). "Verbal correlates of perceptual dimensions of timbre." In: *International Conference on Music Perception and Cognition (ICMPC)* (cit. on p. 43).

Fenet, S, G Richard, and Y Grenier (2011). "A scalable audio fingerprint method with robustness to pitch-shifting." In: *Proceedings of the International Society for Music Information Retrieval (ISMIR)* (cit. on p. 24).

Flanagan, JL (1960). "Models for approximating basilar membrane displacement." In: *Journal of the Acoustical Society of America* 32.7, pp. 937–937 (cit. on p. 28).

Flanagan, JL and RM Golden (1966). "Phase vocoder." In: *Bell System Technical Journal* 45.9, pp. 1493–1509 (cit. on p. 32).

Flandrin, P (1998). *Time-frequency/time-scale analysis*. Academic press (cit. on pp. 1, 64, 117).

— (2001). "Time-frequency and chirps." In: vol. 4391, pp. 161–175 (cit. on pp. 25, 26).

Flandrin, P, F Auger, E Chassande-Mottin, et al. (2002). "Time-Frequency reassignment from principles to algorithms." In: *Applications in time-frequency signal processing* 5, pp. 179–203 (cit. on p. 64).

Fleet, D and Y Weiss (2006). "Optical flow estimation." In: *Handbook of mathematical models in computer vision*. Springer, pp. 237–257 (cit. on p. 119).

Fletcher, NH and T Rossing (2012). *The physics of musical instruments*. Springer Science & Business Media (cit. on p. 1).

Font, F, G Roma, and X Serra (2013). "Freesound technical demo." In: *Proceedings of the ACM International Conference on Multimedia* (cit. on p. 82).

Fourer, D et al. (2014). "Automatic Timbre Classification of Ethnomusicological Audio Recordings." In: *International Society for Music Information Retrieval Conference (ISMIR)* (cit. on p. 133).

Fousek, P, P Dognin, and V Goel (2015). "Evaluating deep scattering spectra with deep neural networks on large-scale spontaneous speech task." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 54).

Fritsch, J and MD Plumbley (2013). "Score-informed audio source separation using constrained nonnegative matrix factorization and score synthesis." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 137).

Gillet, Olivier and Gaël Richard (2004). "Automatic transcription of drum loops." In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (cit. on p. 49).

Giryes, R, G Sapiro, and AM Bronstein (2015). "Deep neural networks with random Gaussian weights: a universal classification strategy?" In: *IEEE Transactions on Signal Processing* 64.13, pp. 3444–3457 (cit. on p. 145).

Glasberg, BR and BCJ Moore (1990). "Derivation of auditory filter shapes from notched-noise data." In: *Hearing research* 47.1-2, pp. 103–138 (cit. on p. 17).

Gold, B, N Morgan, and D Ellis (2011). *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons (cit. on p. 107).

Goto, M et al. (2002). "RWC Music Database: Popular, Classical and Jazz Music Databases." In: *Proceedings of the International Society in Music Information Retrieval (ISMIR)* (cit. on p. 132).

— (2003). "RWC Music Database: Music genre database and musical instrument sound database." In: *Proceedings of the International Society in Music Information Retrieval (ISMIR)* (cit. on pp. 42, 98, 135).

Grey, JM and JW Gordon (1978). "Perceptual effects of spectral modifications on musical timbres." In: *Journal of the Acoustical Society of America* 63.5, pp. 1493–1500 (cit. on p. 14).

Gribonval, Rémi (2001). "Fast matching pursuit with a multiscale dictionary of Gaussian chirps." In: *IEEE Transactions on Signal Processing* 49.5, pp. 994–1001 (cit. on p. 64).

Grossmann, A and J Morlet (1984). "Decomposition of Hardy functions into square integrable wavelets of constant shape." In: *SIAM Journal on Mathematical Analysis* 15.4, pp. 723–736 (cit. on p. 29).

Hamel, P, Y Bengio, and D Eck (2012). "Building musically-relevant audio features through multiple timescale representations." In: *Proceedings of the International Conference on Music Information Retrieval (ISMIR)* (cit. on p. 148).

Handel, S and ML Erickson (2001). "A rule of thumb: The bandwidth for timbre invariance is one octave." In: *Music Perception: An Interdisciplinary Journal* 19.1, pp. 121–126 (cit. on p. 44).

Hemery, E and JJ Aucouturier (2015). "One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis." In: *Frontiers in computational neuroscience* 9 (cit. on pp. 38, 66).

Herrera-Boyer, P, G Peeters, and S Dubnov (2003). "Automatic classification of musical instrument sounds." In: *Journal of New Music Research* 32.1, pp. 3–21 (cit. on p. 135).

Hsu, CW and CJ Lin (2002). "A comparison of methods for multiclass support vector machines." In: *IEEE Transactions on Neural Networks* 13.2, pp. 415–425 (cit. on p. 88).

Humphrey, EJ, JP Bello, and Y Le Cun (2013). "Feature learning and deep architectures: new directions for music informatics." In: *Journal of Intelligent Information Systems* 41.3, pp. 461–481 (cit. on pp. 8, 145).

Humphrey, EJ, T Cho, and JP Bello (2012). "Learning a robust Tonnetz-space transform for automatic chord recognition." In: *Proceedings of the IEEE International Conference on Acoustics, Sound and Signal Conference (ICASSP)* (cit. on p. 141).

Jedrzejewski, F (2002). *Mathématiques des systèmes acoustiques: tempéraments et modèles contemporains*. L'Harmattan (cit. on pp. 16, 101).

Jiang, N et al. (2011). "Analyzing chroma feature types for automated chord recognition." In: *Proceedings of the Audio Engineering Society Conference (AES)* (cit. on p. 107).

Joder, Cyril, Slim Essid, and Gaël Richard (2009). "Temporal integration for audio classification with application to musical instrument classification." In: *IEEE Transactions on Audio, Speech and Language Processing* 17.1, pp. 174–186 (cit. on pp. 8, 21, 88, 133, 136, 137, 140).

Kaminskyj, I and T Czaszejko (2005). "Automatic recognition of isolated monophonic musical instrument sounds using kNNC." In: *Journal of Intelligent Information Systems* 24.2-3, pp. 199–221 (cit. on p. 135).

Kanedera, N et al. (1999). "On the relative importance of various components of the modulation spectrum for automatic speech recognition." In: *Speech Communication* 28.1, pp. 43–55 (cit. on p. 53).

Kereliuk, C and P Depalle (2008). "Improved hidden Markov model partial tracking through time-frequency analysis." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 32).

Kereliuk, C, BL Sturm, and J Larsen (2015). "Deep Learning and Music Adversaries." In: *IEEE Transactions on Multimedia* 17.11, pp. 2059–2071 (cit. on p. 141).

Kim, HS and JO Smith (2014). "Short-time time reversal on audio signals." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 27).

— (2015). "Harmonizing effect using short-time time reversal." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 27).

Klapuri, AP (2003). "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness." In: *IEEE Transactions on Speech and Audio Processing* 11.6, pp. 804–816 (cit. on pp. 119, 120).

Klein, DJ et al. (2000). "Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design." In: *Journal of Computational Neuroscience* 9.1, pp. 85–111 (cit. on p. 65).

Kleinschmidt, M (2002). "Methods for capturing spectro-temporal modulations in automatic speech recognition." In: *Acta Acustica* 88.3, pp. 416–422 (cit. on p. 66).

Kolozali, S et al. (2011). "Knowledge Representation Issues in Musical Instrument Ontology Design." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on p. 138).

Kostka, S (2016). *Materials and Techniques of Post Tonal Music*. Routledge (cit. on p. 136).

Krishna, AG and TV Sreenivas. "Music instrument recognition: from isolated notes to solo phrases." In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 265–269 (cit. on p. 136).

Kronland-Martinet, R (1988). "The wavelet transform for analysis, synthesis, and processing of speech and music sounds." In: *Computer Music Journal* 12.4, pp. 11–20 (cit. on pp. 32, 106).

Kuhn, TS (1962). *The Structure of Scientific Revolutions*. Ed. by University of Chicago Press (cit. on p. 149).

Lagrange, M, S Marchand, and JB Rault (2007). "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds." In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.5, pp. 1625–1634 (cit. on p. 32).

Le Cun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning." In: *Nature* 521.7553, pp. 436–444 (cit. on pp. 81, 149).

Le Roux, N et al. (2007). "Learning the 2-D Topology of Images." In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1–8 (cit. on p. 97).

Lei, H, BT Meyer, and N Mirghafori (2012). "Spectro-temporal Gabor features for speaker recognition." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 66).

Li, P, J Qian, and T Wang (2015). "Automatic Instrument Recognition in Polyphonic Music Using Convolutional Neural Networks." In: *arXiv preprint 1511.05520* (cit. on pp. 133, 141, 149).

Lidy, T and A Schindler (2016). "CQT-based convolutional neural networks for audio scene classification." In: *Proceedings of the IEEE AASP Workshop on Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)* (cit. on p. 81).

Lindeberg, T and A Friberg (2015). "Idealized computational models for auditory receptive fields." In: *PLoS one* 10.3 (cit. on p. 66).

Liuni, M and A Röbel (2013). "Phase vocoder and beyond." In: *Musica/Tecnologia* 7, pp. 73–89 (cit. on p. 32).

Livshin, A and X Rodet (2003). "The importance of cross-database evaluation in sound classification." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on p. 137).

— (2004). "Musical instrument identification in continuous recordings." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 136).
Lostanlen, V and CE Cella (2016). "Deep convolutional networks on the pitch spiral for musical instrument recognition." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on pp. 11, 93, 141, 142, 145, 147).
Lostanlen, V and S Mallat (2015). "Wavelet Scattering on the Pitch Spiral." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)*. Vol. 18, pp. 429–432 (cit. on p. 93).
Lyon, RF, AG Katsiamis, and EM Drakakis (2010). "History and future of auditory filter models." In: *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCS)* (cit. on p. 28).
Malik, J and R Rosenholtz (1997). "Computing local surface orientation and shape from texture for curved surfaces." In: *International Journal of Computer Vision* 23.2, pp. 149–168 (cit. on p. 27).
Mallat, S (2008). *A wavelet tour of signal processing: the sparse way*. Academic press (cit. on pp. ix, 12, 15, 25, 31, 34, 41).
— (2012). "Group invariant scattering." In: *Communications on Pure and Applied Mathematics* 65.10, pp. 1331–1398 (cit. on pp. 56, 145).
— (2016). "Understanding deep convolutional networks." In: *Philosophical Transactions of the Royal Society A* 374.2065 (cit. on pp. 145, 146).
Mallat, S and I Waldspurger (2015). "Phase retrieval for the Cauchy wavelet transform." In: *Journal of Fourier Analysis and Applications* 21.6, pp. 1251–1309 (cit. on p. 70).
Mallat, SG (1989). "A theory for multiresolution signal decomposition: the wavelet representation." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.7, pp. 674–693 (cit. on p. 55).
Mallat, SG and Z Zhang (1993). "Matching pursuits with time-frequency dictionaries." In: *IEEE Transactions on Signal Processing* 41.12, pp. 3397–3415 (cit. on pp. 64, 80).
Mallat, Stephane and Wen Liang Hwang (1992). "Singularity detection and processing with wavelets." In: *IEEE transactions on information theory* 38.2, pp. 617–643 (cit. on p. 102).
Malm, WP (2000). *Traditional Japanese Music and Musical Instruments*. Kodansha International (cit. on p. 95).
Mann, S and S Haykin (1995). "The chirplet transform: physical considerations." In: *IEEE Transactions on Signal Processing* 43.11, pp. 2745–2761 (cit. on p. 64).
Martin, KD and YE Kim (1998). "Musical instrument identification: a pattern recognition approach." In: *Proceedings of the Acoustical Society of America* (cit. on p. 135).
Mauch, M and S Dixon (2014). "pYIN: A fundamental frequency estimator using probabilistic threshold distributions." In: *Proceedings*

of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (cit. on p. 37).

Mauch, M, H Fujihara, and M Goto (2012). "Integrating Additional Chord Information Into HMM-Based Lyrics-to-Audio Alignment." In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.1, pp. 200–210 (cit. on p. 106).

McAdams, S et al. (1995). "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes." In: *Psychological research* 58.3, pp. 177–192 (cit. on p. 43).

McAulay, R and T Quatieri (1986). "Speech analysis/synthesis based on a sinusoidal representation." In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34.4, pp. 744–754 (cit. on p. 32).

McDermott, JH and EP Simoncelli (2011). "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis." In: *Neuron* 71.5, pp. 926–940 (cit. on pp. 4, 5, 70, 73, 146).

McFee, B, EJ Humphrey, and JP Bello (2015). "A software framework for musical data augmentation." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on pp. 133, 141, 149).

McFee, B, EJ Humphrey, and J Urbano (2016). "A Plan for Sustainable MIR Evaluation." In: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (cit. on p. 149).

McKinney, MF and J Breebaart (2003). "Features for audio and music classification." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on p. 53).

McVicar, M et al. (2014). "Automatic chord estimation from audio: a review of the state of the art." In: *IEEE Transactions on Audio, Speech, and Language Processing* 22.2, pp. 556–575 (cit. on pp. 21, 106, 107).

Mesaros, A, T Heittola, and T Virtanen (2016). "TUT database for acoustic scene classification and sound event detection." In: *Proceedings of the European Signal Processing Conference (EUSIPCO)* (cit. on p. 90).

Mesgarani, N, M Slaney, and SA Shamma (2006). "Discrimination of speech from nonspeech based on multiscale spectrotemporal modulations." In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.3, pp. 920–930 (cit. on p. 66).

Meyer, LB (1989). *Style and Music: Theory, History, and Ideology*. University of Chicago Press (cit. on p. 39).

Milner, Robin (1978). "A theory of type polymorphism in programming." In: *Journal of computer and system sciences* 17.3, pp. 348–375 (cit. on p. 61).

Mitra, V et al. (2012). "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 49).

Mohajer, K et al. (2010). *System and method for storing and retrieving non-text-based information*. US Patent 7,788,279 (cit. on p. 54).

Müller, M (2007). *Information retrieval for music and motion*. Springer (cit. on p. 34).

Müller, M, S Ewert, and S Kreuzer (2009). "Making chroma features more robust to timbre changes." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (cit. on p. 107).

Mydlarz, C, J Salamon, and JP Bello (2017). "The implementation of low-cost urban acoustic monitoring devices." In: *Applied Acoustics* 117, pp. 207–218 (cit. on p. 148).

Necciari, T et al. (2013). "The ERBlet transform: an auditory-based time-frequency representation with perfect reconstruction." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on p. 18).

Nelken, I and A de Cheveigné (2013). "An ear for statistics." In: *Nature Neuroscience* 16.4, pp. 381–382 (cit. on p. 80).

Nettl, B (1956). *Music in primitive culture*. Harvard University Press (cit. on pp. 95, 147).

Noll, AM and MR Schroeder (1967). "Short-Time cepstrum pitch detection." In: *Journal of the Acoustical Society of America* 36.5, pp. 1030–1030 (cit. on p. 37).

Norman, L et al. (2010). "Player control of brassiness at intermediate dynamic levels in brass instruments." In: *Acta Acustica* 96.4, pp. 614–621 (cit. on p. 44).

Omer, H and B Torrésani (2016). "Time-frequency and time-scale analysis of deformed stationary processes, with application to non-stationary sound modeling." In: *Applied and Computational Harmonic Analysis* (cit. on pp. 33, 119).

Oord, A van den et al. (2016). "Wavenet: a generative model for raw audio." In: *arXiv preprint 1609.03499* (cit. on p. 148).

Oord, Aaron van den, Sander Dieleman, and Benjamin Schrauwen (2013). "Deep content-based music recommendation." In: *Advances in Neural Information Processing Systems (NIPS)* (cit. on p. 148).

Oppenheim, AV and RW Schafer (2004). "From frequency to quefrency: a history of the cepstrum." In: *IEEE Signal Processing Magazine* 21.5, pp. 95–106 (cit. on p. 36).

Oyallon, E and S Mallat (2015). "Deep roto-translation scattering for object classification." In: (cit. on p. 145).

Patil, K and M Elhilali (2015). "Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases." In: *EURASIP Journal on Audio, Speech, and Music Processing* (cit. on pp. 66, 135–137).

Patil, K et al. (2012). "Music in our ears: the biological bases of musical timbre perception." In: *PLoS Computational Biology* 8.11 (cit. on pp. 66, 135, 146).

Patterson, RD (1976). "Auditory filter shapes derived with noise stimuli." In: *Journal of the Acoustical Society of America* 59.3, pp. 640–654 (cit. on p. 28).

Paulus, J, M Müller, and A Klapuri (2010). "Audio-based music structure analysis." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on p. 107).

Peddinti, V et al. (2014). "Deep scattering spectrum with deep neural networks." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 55, 148).

Peeters, G (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Tech. rep. Ircam (cit. on p. 36).

— (2006). "Chroma-based estimation of musical key from audio-signal analysis." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on p. 107).

Peeters, G, A La Burthe, and X Rodet (2002). "Toward automatic music audio summary generation from signal analysis." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on p. 49).

Peeters, G and X Rodet (2003). "Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instruments databases." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 138).

Pelofi, C et al. (2017). "Interindividual variability in auditory scene analysis revealed by confidence judgements." In: *Proceedings of the Philosophical Transactions of the Royal Society B* 372.1714 (cit. on p. 103).

Pérez-Sancho, C, D Rizo, and JM Inesta (2009). "Genre classification using chords and stochastic language models." In: *Connection science* 21.2-3, pp. 145–159 (cit. on p. 106).

Pesch, L (2009). *The Oxford illustrated companion to South Indian classical music*. Oxford University Press (cit. on p. 95).

Piczak, KJ (2015a). "Environmental sound classification with convolutional neural networks." In: *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (cit. on pp. 81, 89).

— (2015b). "ESC: dataset for environmental sound classification." In: *Proceedings of the ACM International Conference on Multimedia* (cit. on pp. 8, 81, 82).

Polfreman, R (2013). "Comparing onset detection and perceptual attack time." In: *Proceedings of the International Society for Music Information Retrieval (ISMIR)* (cit. on p. 21).

Pressnitzer, D and D Gnansia (2005). "Real-time auditory models." In: *Proceedings of the International Computer Music Conference (ICMC)* (cit. on p. 28).

Qiu, A, CE Schreiner, and MA Escabí (2003). "Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition." In: *Journal of Neurophysiology* 90.1, pp. 456–476 (cit. on p. 66).

Risset, JC (1965). "Computer study of trumpet tones." In: *Journal of the Acoustical Society of America* 38.5, pp. 912–912 (cit. on p. 31).

— (1969). "Pitch control and pitch paradoxes demonstrated with computer-synthesized sounds." In: *Journal of the Acoustical Society of America* 46.1A, pp. 88–88 (cit. on pp. 93, 111).

Risset, JC and MV Mathews (1969). "Analysis of musical-instrument tones." In: *Physics today* 22, p. 23 (cit. on p. 44).

Röbel, A (2003). "A new approach to transient processing in the phase vocoder." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 32).

Röbel, A and X Rodet (2005). "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 42).

Robinson, K and RD Patterson (1995). "The duration required to identify the instrument, the octave, or the pitch chroma of a musical note." In: *Music Perception: An Interdisciplinary Journal* 13.1, pp. 1–15 (cit. on p. 136).

Rodet, X, L Worms, and G Peeters (2003). *Procédé de caractérisation d'un signal sonore*. WO Patent App. PCT/FR2002/004,549 (cit. on p. 54).

Rodríguez-Algarra, F, BL Sturm, and H Maruri-Aguilar (2016). "Analysing scattering-based music content analysis systems: where's the music?" In: *Proceedings of the International Society for Music Information Retrieval conference (ISMIR)* (cit. on p. 54).

Roma, G et al. (2013). "Recurrence quantification analysis features for auditory scene classification." In: *Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)* (cit. on p. 90).

Rowe, R (1992). *Interactive Music Systems: Machine Listening and Composing*. Cambridge, MA, USA: MIT Press (cit. on p. 1).

Rumelhart, DE, GE Hinton, and RJ Williams (1986). "Learning representations by back-propagating errors." In: *Nature* 323.6088, pp. 533–536 (cit. on p. 72).

Sainath, TN et al. (2014). "Deep scattering spectra with deep neural networks for LVCSR tasks." In: *Proceedings of the INTERSPEECH Conference* (cit. on p. 148).

Sainath, TN et al. (2015). "Learning the speech front-end with raw waveform CLDNNs." In: *Proceedings of INTERSPEECH* (cit. on p. 149).

Sakoe, H and S Chiba (1978). "Dynamic programming algorithm optimization for spoken word recognition." In: *IEEE Transactions*

on Acoustics, Speech, and Signal Processing 26.1, pp. 43–49 (cit. on p. 34).

Salamon, J (2016). "Pitch analysis for active music discovery." In: *Machine Learning for Music Discovery workshop, International Conference on Machine Learning (ICML)* (cit. on p. 137).

Salamon, J and JP Bello (2015a). "Feature learning with deep scattering for urban sound analysis." In: *Proceedings of the Signal Processing Conference (EUSIPCO)* (cit. on p. 81).

— (2015b). "Unsupervised feature learning for urban sound classification." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 81, 88).

— (2016). "Deep convolutional neural networks and data augmentation for environmental sound classification." In: *IEEE Signal Processing Letters (accepted)* (cit. on pp. 82, 149).

Salamon, J, C Jacoby, and JP Bello (2014). "A dataset and taxonomy for urban sound research." In: *Proceedings of the IEEE International Conference on Multimedia* (cit. on pp. 1, 8, 82).

Salamon, J et al. (2016). "Towards the automatic classification of avian flight calls for bioacoustic monitoring." In: *PLoS One* 11.11 (cit. on p. 148).

Schädler, MR and B Kollmeier (2015). "Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition." In: *Journal of the Acoustical Society of America* 137.4, pp. 2047–2059 (cit. on p. 67).

Schädler, MR, BT Meyer, and B Kollmeier (2012). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition." In: *Journal of the Acoustical Society of America* 131.5, pp. 4134–4151 (cit. on p. 66).

Schedl, M, E Gómez, and J Urbano (2014). "Music information retrieval: recent developments and applications." In: *Foundations and Trends in Information Retrieval* 8.2–3, pp. 127–261 (cit. on p. 1).

Schlüter, J and S Böck (2014). "Improved musical onset detection with convolutional neural networks." In: *Proceedings of the IEEE International Conference on Acoustics, Sound, and Signal Processing (ICASSP)* (cit. on p. 141).

Schmidhuber, J (2015). "Deep learning in neural networks: an overview." In: *Neural Networks* 61, pp. 85–117 (cit. on p. 149).

Schörkhuber, C and A Klapuri (2010). *Constant-Q transform toolbox for music processing* (cit. on p. 32).

Schörkhuber, C, A Klapuri, and A Sontacchi (2013). "Audio Pitch Shifting Using the Constant-Q Transform." In: *Journal of the Audio Engineering Society* 61.7–8, pp. 562–572 (cit. on p. 32).

Schröder, J et al. (2013). "Acoustic event detection using signal enhancement and spectro-temporal feature extraction." In: *Proceed-*

ings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (cit. on p. 66).

Schroeder, MR (1986). "Auditory paradox based on fractal waveform." In: *Journal of the Acoustical Society of America* 79.1, pp. 186–189 (cit. on p. 102).

Schwarz, D (2011). "State of the art in sound texture synthesis." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (cit. on p. 70).

Sek, A and BCJ Moore (2003). "Testing the concept of a modulation filter bank: the audibility of component modulation and detection of phase change in three-component modulators." In: *Journal of the Acoustical Society of America* 113.5, pp. 2801–2811 (cit. on p. 53).

Seo, JS et al. (2006). "Audio fingerprinting based on normalized spectral subband moments." In: *IEEE Signal Processing Letters* 13.4, pp. 209–212 (cit. on p. 54).

Serra, J et al. (2008). "Chroma binary similarity and local alignment applied to cover song identification." In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.6, pp. 1138–1151 (cit. on p. 107).

Shepard, R (1964). "Circularity in Judgments of Relative Pitch." In: *Journal of the Acoustical Society of America* 36.12, p. 2346 (cit. on pp. 93, 99, 101, 103, 104).

Siedenburg, K, I Fujinaga, and S McAdams (2016). "A comparison of approaches to timbre descriptors in music information retrieval and music psychology." In: *Journal of New Music Research* 45.1, pp. 27–41 (cit. on pp. 43, 66).

Sifre, L and S Mallat (2013). "Rotation, scaling and deformation invariant scattering for texture discrimination." In: *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 145).

Smaragdis, P and JC Brown (2003). "Non-negative matrix factorization for polyphonic music transcription." In: *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* IEEE, pp. 177–180 (cit. on p. 55).

Steele, KM and AK Williams (2006). "Is the bandwidth for timbre invariance only one octave?" In: *Music Perception: An Interdisciplinary Journal* 23.3, pp. 215–220 (cit. on p. 44).

Stevens, S, John Volkmann, and Edwin B Newman (1937). "A scale for the measurement of the psychological magnitude pitch." In: *Journal of the Acoustical Society of America* 8.3, pp. 185–190 (cit. on p. 18).

Stowell, D and MD Plumbley (2014). "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning." In: *PeerJ* 2, e488 (cit. on p. 81).

Stowell, D. et al. (2015). "Detection and Classification of Acoustic Scenes and Events." In: *IEEE Transactions on Multimedia* 17.10, pp. 1733–1746 (cit. on pp. 8, 21, 85, 90).

Stowell, Dan (2015). "Simple-minded audio classifier in Python (using MFCC and GMM)." In: *https://code.soundsoftware.ac.uk/projects/smacpy* (cit. on p. 38).

Sturm, BL (2012). "An analysis of the GTZAN music genre dataset." In: *Proceedings of the International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*. ACM (cit. on p. 136).

— (2014a). "A simple method to determine if a music information retrieval system is a "horse"." In: *IEEE Transactions on Multimedia* 16.6, pp. 1636–1644 (cit. on p. 136).

— (2014b). "The state of the art ten years after a state of the art: future research in music information retrieval." In: *Journal of New Music Research* 43.2, pp. 147–172 (cit. on pp. 21, 53).

Tenenbaum, JB, V De Silva, and JC Langford (2000). "A global geometric framework for nonlinear dimensionality reduction." In: *Science* 290.5500, pp. 2319–2323 (cit. on p. 98).

Theunissen, FE, K Sen, and AJ Doupe (2000). "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds." In: *Journal of Neuroscience* 20.6, pp. 2315–2331 (cit. on p. 65).

Thoret, É et al. (2014). "From sound to shape: auditory perception of drawing movements." In: *Journal of Experimental Psychology: Human Perception and Performance* 40.3, p. 983 (cit. on p. 27).

Tzanetakis, G and P Cook (2002). "Musical genre classification of audio signals." In: *IEEE Transactions on Speech and Audio Processing* 10.5, pp. 293–302 (cit. on p. 53).

Ullrich, K, J Schlüter, and T Grill (2014). "Boundary detection in music structure analysis using convolutional neural networks." In: *Proceedings of the International Society on Music Information Retrieval (ISMIR)* (cit. on p. 141).

Umesh, S, L Cohen, and D Nelson (1999). "Fitting the Mel scale." In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (cit. on p. 18).

Valenti, M et al. (2016). "Acoustic Scene Classification Using Convolutional Neural Networks." In: *Proceedings of the IEEE AASP Workshop on Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE)* (cit. on p. 81).

Van Aalst, Jules A (2012). *Chinese music*. Cambridge University Press (cit. on p. 95).

Van Loan, Charles (1992). *Computational Frameworks for the Fast Fourier Transform*. Society for Industrial and Applied Mathematics (cit. on p. 16).

Venkitaraman, A, A Adiga, and CS Seelamantula (2014). "Auditory-motivated Gammatone wavelet transform." In: *Signal Processing* 94, pp. 608–619 (cit. on p. 29).

Versteegh, M et al. (2015). "The zero-resource speech challenge 2015." In: *Proceedings of INTERSPEECH* (cit. on p. 55).

Vinet, Hugues (2003). "The representation levels of music information." In: *Proceedings of the International Symposium on Computer Music Modeling and Retrieval (CMMR)*. Springer (cit. on p. 21).

Wakefield, GH (1999). "Mathematical representation of joint time-chroma distributions." In: *Proceedings of the SPIE International Symposium on Optical Science, Engineering, and Instrumentation*. International Society for Optics and Photonics (cit. on p. 106).

Wang, Avery (2006). "The Shazam Music Recognition Service." In: *Communications of the ACM* 49.8, pp. 44–48 (cit. on p. 54).

Wang, D and GJ Brown (2006). *Computational auditory scene analysis: principles, algorithms, and applications*. Wiley-IEEE Press (cit. on p. 85).

Warren, JD et al. (2003a). "Analyzing pitch chroma and pitch height in the human brain." In: *Annals of the New York Academy of Sciences* 999.1, pp. 212–214 (cit. on p. 95).

— (2003b). "Separating pitch chroma and pitch height in the human brain." In: *Proceedings of the National Academy of Sciences* 100.17, pp. 10038–10042 (cit. on pp. 95, 97).

Wieczorkowska, AA and JM Żytkow (2003). "Analysis of feature dependencies in sound description." In: *Journal of Intelligent Information Systems* 20.3, pp. 285–302 (cit. on p. 135).

Wishart, T (1988). "The Composition of "Vox-5"." In: *Computer Music Journal* 12.4, pp. 21–27 (cit. on p. 32).

Wuerger, S and R Shapley (1996). "On the visually perceived direction of motion by Hans Wallach: 60 years later." In: *Perception* 25.11, pp. 1317–1368 (cit. on p. 111).

Zeghidour, N et al. (2016). "A deep scattering spectrum: deep Siamese network pipeline for unsupervised acoustic modeling." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cit. on pp. 55, 148).

Résumé

Dans le cadre de la classification de sons, cette thèse construit des représentations du signal qui vérifient des propriétés d'invariance et de variabilité inter-classe. D'abord, nous étudions le scattering temps-fréquence, une représentation qui extrait des modulations spectrotemporelles à différentes échelles. En classification de sons urbains et environnementaux, nous obtenons de meilleurs résultats que les réseaux profonds à convolutions et les descripteurs à court terme. Ensuite, nous introduisons le scattering en spirale, une représentation qui combine des transformées en ondelettes selon le temps, selon les log-fréquences, et à travers les octaves. Le scattering en spirale suit la géométrie de la spirale de Shepard, qui fait un tour complet à chaque octave. Nous étudions les sons voisés avec un modèle source-filtre non stationnaire dans lequel la source et le filtre sont transposés au cours du temps, et montrons que le scattering en spirale sépare et linéarise ces transpositions. Le scattering en spirale améliore les performances de l'état de l'art en classification d'instruments de musique. Outre la classification de sons, le scattering temps-fréquence et le scattering en spirale peuvent être utilisés comme des descripteurs pour la synthèse de textures audio. Contrairement au scattering temporel, le scattering temps-fréquence est capable de capturer la cohérence de motifs spectrotemporels en bioacoustique et en parole, jusqu'à une échelle d'intégration de 500 ms environ. À partir de ce cadre d'analyse-synthèse, une collaboration art-science avec le compositeur Florian Hecker a mené à la création de cinq œuvres musicales.

Mots Clés

Classification de sons, transformée en scattering, synthèse de textures auditives, spirale de Shepard, musique par ordinateur.

Abstract

This dissertation addresses audio classification by designing signal representations which satisfy appropriate invariants while preserving inter-class variability. First, we study time-frequency scattering, a representation which extract modulations at various scales and rates in a similar way to idealized models of spectrotemporal receptive fields in auditory neuroscience. We report state-of-the-art results in the classification of urban and environmental sounds, thus outperforming short-term audio descriptors and deep convolutional networks. Secondly, we introduce spiral scattering, a representation which combines wavelet convolutions along time, along log-frequency, and across octaves. Spiral scattering follows the geometry of the Shepard pitch spiral, which makes a full turn at every octave. We study voiced sounds with a nonstationary source-filter model where both the source and the filter are transposed through time, and show that spiral scattering disentangles and linearizes these transpositions. Furthermore, spiral scattering reaches state-of-the-art results in musical instrument classification of solo recordings. Aside from audio classification, time-frequency scattering and spiral scattering can be used as summary statistics for audio texture synthesis. We find that, unlike the previously existing temporal scattering transform, time-frequency scattering is able to capture the coherence of spectrotemporal patterns, such as those arising in bioacoustics or speech, up to an integration scale of about 500 ms. Based on this analysis-synthesis framework, an artistic collaboration with composer Florian Hecker has led to the creation of five computer music pieces.

Keywords

Audio classification, scattering transform, audio texture synthesis, spirale de Shaprd, computer music.