

Deep Scattering Spectrum

Joakim Andén, Stéphane Mallat*

Abstract

A scattering transform defines a locally translation invariant representation which is stable to time-warping deformations. It extends MFCC representations by computing modulation spectrum coefficients of multiple orders, through cascades of wavelet convolutions and modulus operators. Second-order scattering coefficients characterize transient phenomena such as attacks and amplitude modulation. A frequency transposition invariant representation is obtained by applying a scattering transform along log-frequency. State-the-of-art classification results are obtained for musical genre and phone classification on GTZAN and TIMIT databases, respectively.

Keywords: Audio classification, deep neural networks, MFCC, modulation spectrum, wavelets.

1 Introduction

A major difficulty of audio representations for classification is the multiplicity of information at different time scales: pitch and timbre at the scale of milliseconds, the rhythm of speech and music at the scale of seconds, and the music progression over minutes and hours. Mel-frequency cepstral coefficients (MFCCs) are efficient local descriptors at time scales up to 25ms. Capturing larger structures up to 500ms is however necessary in most applications. This paper studies the construction of stable, invariant signal representations over such larger time scales. We concentrate on audio applications, but introduce a generic scattering representation for classification, which applies to many signal modalities beyond audio [12].

Spectrograms compute locally invariant descriptors over time intervals limited by a window. Section 2 shows that high-frequency spectrogram coefficients are not stable to variability due to time-warping deformations, which occur in most signals, particularly in audio. MFCCs average spectrogram values over mel-frequency bands, which improves stability to time warping but also removes information. Over time intervals larger than 25ms, the information loss becomes too important, which is why MFCCs are limited to such short time intervals.

*This work is supported by the ANR 10-BLAN-0126 and ERC InvariantClass 320959 grants.

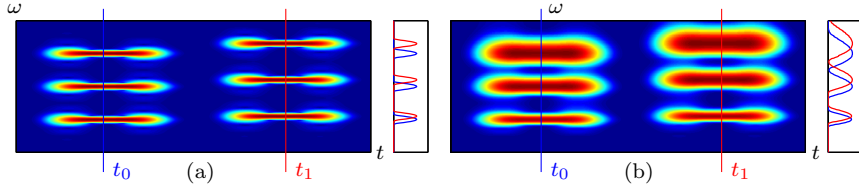


Figure 1: (a) Spectrogram $\log |\widehat{x}(t, \omega)|$ for a harmonic signal $x(t)$ (centered in t_0) followed by $\log |\widehat{x}_\tau(t, \omega)|$ for $x_\tau(t) = x((1 - \epsilon)t)$ (centered in t_1), as a function of t and ω . The right graph plots $\log |\widehat{x}(t_0, \omega)|$ (blue) and $\log |\widehat{x}_\tau(t_1, \omega)|$ (red) as a function of ω . Their partials do not overlap at high frequencies. (b) Mel-frequency spectrogram $\log Mx(t, \omega)$ followed by $\log Mx_\tau(t, \omega)$. The right graph plots $\log Mx(t_0, \omega)$ (blue) and $\log Mx_\tau(t_1, \omega)$ (red) as a function of ω . With a mel-scale frequency averaging, the partials of x and x_τ overlap at all frequencies.

Modulation spectrum decompositions [2, 17, 23, 26, 32, 36, 37, 40, 42] characterize the temporal evolution of mel-frequency spectrograms over larger time scales, with autocorrelation or Fourier coefficients. However, this modulation spectrum also suffers from instability to time-warping deformation, which impedes classification performance.

Section 3 shows that the information lost by mel-frequency spectral coefficients can be recovered with multiple layers of wavelet coefficients, which are stable to time-warping deformations. A scattering transform [31] computes such a cascade of wavelet transforms and modulus non-linearities. Its computational structure is similar to a convolutional deep neural network [3, 14, 21, 24, 25, 27, 34], but involves no learning. It outputs time-averaged coefficients, providing informative signal invariants over potentially large time scales.

A scattering transform has striking similarities with physiological models of the cochlea and of the auditory pathway [11, 15], also used for audio processing [33]. Its energy conservation and other mathematical properties are reviewed in Section 4. An approximate inverse scattering transform is introduced in Section 5, with numerical examples. Section 6 relates the amplitude of scattering coefficients to audio signal properties. These coefficients provide accurate measurements of frequency intervals between harmonics and also characterize the amplitude modulation of voiced and unvoiced sounds. The logarithm of scattering coefficients linearly separates audio components related to pitch, formant and timbre.

Frequency transpositions form another important source of audio variability, which should be kept or removed depending upon the classification task. For example, speaker-independent phone recognition requires some frequency transposition invariance, while frequency localization is necessary for speaker identification. Section 7 shows that cascading a scattering transform along log-frequency yields a transposition invariant representation which is stable to frequency deformation.

Scattering representations have proved useful for image [5, 39] and audio [1, 4, 10] classification. Section 8 explains how to adapt and optimize the amount of time and frequency invariance for each signal class, at the supervised learning stage. A time and frequency scattering representation is used for musical genre classification over the GTZAN database, and for phone classification over the TIMIT corpus. State-of-the-art results are obtained with a Gaussian kernel SVM applied to scattering feature vectors. All figures and results are reproducible using a MATLAB software package, available at <http://www.di.ens.fr/signal/scattering/>.

2 Mel-frequency Spectrum

Section 2.1 shows that high-frequency spectrogram coefficients are not stable to time-warping deformation. The mel-frequency spectrogram stabilizes these coefficients by averaging them along frequency, but loses information. To analyze this information loss, Section 2.2 relates the mel-frequency spectrogram to the amplitude output of a filter bank which computes a wavelet transform.

2.1 Fourier Invariance and Deformation Instability

Let $\hat{x}(\omega) = \int x(u) e^{-i\omega u} du$ be the Fourier transform of x . If $x_c(t) = x(t - c)$ then $\hat{x}_c(\omega) = e^{-ic\omega} \hat{x}(\omega)$. The Fourier transform modulus is thus invariant to translation:

$$|\hat{x}_c(\omega)| = |\hat{x}(\omega)|. \quad (1)$$

A spectrogram localizes this translation invariance with a window ϕ of duration T such that $\int \phi(u) du = 1$. It is defined by

$$|\hat{x}(t, \omega)| = \left| \int x(u) \phi(u - t) e^{-i\omega u} du \right|. \quad (2)$$

If $|c| \ll T$ then one can verify that $|\hat{x}_c(t, \omega)| \approx |\hat{x}(t, \omega)|$.

Suppose that x is not just translated but time-warped to give $x_\tau(t) = x(t - \tau(t))$ with $|\tau'(t)| < 1$. A representation $\Phi(x)$ is said to be stable to deformation if its Euclidean norm $\|\Phi(x) - \Phi(x_\tau)\|$ is small when the deformation is small. The deformation size is measured by $\sup_t |\tau'(t)|$. If it vanishes then it is a “pure” translation without deformation. Stability is formally defined as a Lipschitz continuity condition relatively to this metric. It means that there exists $C > 0$ such that for all τ with $\sup_t |\tau'(t)| < 1$

$$\|\Phi(x) - \Phi(x_\tau)\| \leq C \sup_t |\tau'(t)| \|x\|. \quad (3)$$

A Fourier modulus representation $\Phi(x) = |\hat{x}|$ is not stable to deformation because high frequencies are severely distorted by small deformations. For example, let us consider a small dilation $\tau(t) = \epsilon t$ with $0 < \epsilon \ll 1$. Since $\tau'(t) = \epsilon$, the Lipschitz continuity condition (3) becomes

$$\| |\hat{x}| - |\hat{x}_\tau| \| \leq C \epsilon \|x\|. \quad (4)$$

The Fourier transform of $x_\tau(t) = x((1-\epsilon)t)$ is $\widehat{x}_\tau(\omega) = (1-\epsilon)^{-1} \widehat{x}((1-\epsilon)^{-1}\omega)$. This dilation shifts a frequency ω_0 by $\epsilon|\omega_0|$. For a harmonic signal $x(t) = g(t) \sum_n a_n \cos(n\xi t)$, the Fourier transform is a sum of partials

$$\widehat{x}(\omega) = \sum_n \frac{a_n}{2} \left(\widehat{g}(\omega - n\xi) + \widehat{g}(\omega + n\xi) \right). \quad (5)$$

After time-warping, each partial $\widehat{g}(\omega \pm n\xi)$ is translated by $\epsilon n|\xi|$, as shown in the spectrogram of Figure 1(a). Even though ϵ is small, at high frequencies $n\epsilon$ becomes larger than the bandwidth of \widehat{g} . Consequently, the supports of deformed harmonics $\widehat{g}(\omega(1-\epsilon)^{-1} - n\xi)$ do not overlap those of the original harmonics $\widehat{g}(\omega - n\xi)$ and hence induce a large Euclidean distance. The Euclidean distance of $|\widehat{x}|$ and $|\widehat{x}_\tau|$ thus does not decrease proportionally to ϵ if the harmonic amplitudes a_n are sufficiently large at high frequencies. This proves that the continuity condition (4) is not satisfied.

The autocorrelation $Rx(u) = \int x(t) x^*(t-u) dt$ is also a translation invariant representation which has the same deformation instability as the Fourier transform modulus. Indeed, $\widehat{Rx}(\omega) = |\widehat{x}(\omega)|^2$ so $\|Rx - Rx_\tau\| = (2\pi)^{-1} \| |\widehat{x}|^2 - |\widehat{x}_\tau|^2 \|$.

2.2 Mel-frequency Deformation Stability and Filter Banks

A mel-frequency spectrogram averages the spectrogram energy with mel-scale filters $\widehat{\psi}_\lambda$, where λ is the center frequency of each $\widehat{\psi}_\lambda(\omega)$:

$$Mx(t, \lambda) = \frac{1}{2\pi} \int |\widehat{x}(t, \omega)|^2 |\widehat{\psi}_\lambda(\omega)|^2 d\omega. \quad (6)$$

The band-pass filters $\widehat{\psi}_\lambda$ have a constant-Q frequency bandwidth, with a support centered at λ whose size is proportional to λ . At the lowest frequencies, instead of being constant-Q, the bandwidth of $\widehat{\psi}_\lambda$ remains equal to $2\pi/T$ so that $\psi_\lambda(t)$ is mostly localized in a time interval of size T .

The mel-frequency averaging removes deformation instability. Indeed, larger displacements of high frequencies are compensated by the wider averaging by the kernel $|\widehat{\psi}_\lambda(\omega)|^2$. After this averaging, Figure 1(b) shows that the partials of a harmonic signal x and the partials of its dilation x_τ still overlap at high frequencies. As opposed to spectrograms, a mel-frequency representation $\Phi(x) = Mx$ is Lipschitz stable to deformations in the sense of (3).

This time-warping stability is due to the mel-scale averaging. However, this averaging loses information. We show that this frequency averaging can be rewritten as a time averaging of a filter bank output. Since $\widehat{x}(t, \omega)$ in (2) is the Fourier transform of $x_t(u) = x(u)\phi(u-t)$, applying Plancherel's formula gives

$$Mx(t, \lambda) = \frac{1}{2\pi} \int |\widehat{x}_t(\omega)|^2 |\widehat{\psi}_\lambda(\omega)|^2 d\omega \quad (7)$$

$$= \int |x_t \star \psi_\lambda(v)|^2 dv \quad (8)$$

$$= \int \left| \int x(u)\phi(u-t)\psi_\lambda(v-u)du \right|^2 dv \quad (9)$$

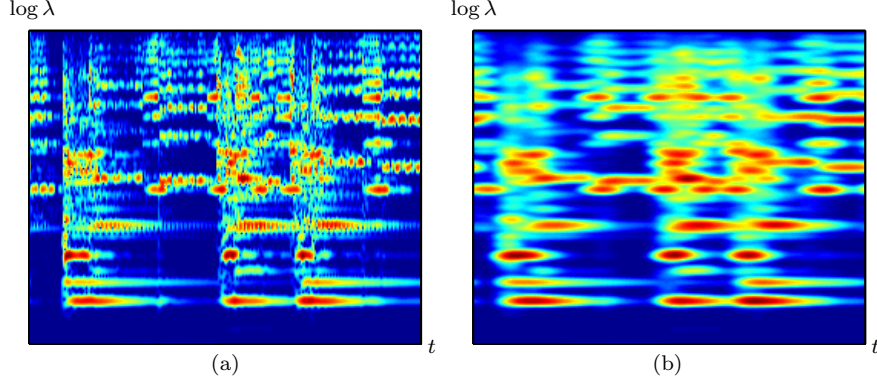


Figure 2: (a): Scalogram $\log |x \star \psi_\lambda(t)|^2$ for a musical signal, as a function of t and λ . (b): Averaged scalogram $\log |x \star \psi_\lambda|^2 \star \phi^2(t)$ with a lowpass filter ϕ of duration $T = 190\text{ms}$.

If $\lambda \gg Q_1/T$ then $\phi(t)$ is approximately constant on the support of $\psi_\lambda(t)$, so $\phi(u-t)\psi_\lambda(v-u) \approx \phi(v-t)\psi_\lambda(v-u)$, and hence

$$Mx(t, \lambda) \approx \int \left| \int x(u) \psi_\lambda(v-u) du \right|^2 |\phi(v-t)|^2 dv \quad (10)$$

$$= |x \star \psi_\lambda|^2 \star |\phi|^2(t). \quad (11)$$

The frequency averaging of the spectrogram is thus nearly equal to the time averaging of $|x \star \psi_\lambda|^2$. Section 3.1 studies the properties of the constant-Q filter bank $\{\psi_\lambda\}_\lambda$, which defines an analytic wavelet transform.

Figures 2(a) and 2(b) display $|x \star \psi_\lambda|^2$ and $|x \star \psi_\lambda|^2 \star |\phi|^2$, respectively, for a musical recording. The window duration is $T = 190\text{ms}$. This time averaging removes fine-scale information such as vibratos and attacks. To reduce information loss, a mel-frequency spectrogram is often computed over small time windows of about 25ms. As a result, it does not capture large-scale structures, which limits classification performance.

To increase T without losing too much information, it is necessary to capture the high frequencies of $|x \star \psi_\lambda|$. This can be done with a modulation spectrum. The modulation spectrum can be defined as the spectrogram [2, 23, 32, 37] of $|x \star \psi_\lambda|$, or as its short-time autocorrelation [36, 40]. However, these modulation spectra are unstable to time-warping deformations. Indeed, a time-warping of x induces a time-warping of $|x \star \psi_\lambda|$, and Section 2.1 showed that spectrograms and autocorrelations have deformation instabilities. Constant-Q averaged modulation spectra [17, 42] stabilize spectrogram representations with another averaging along modulation frequencies. According to (11), this can also be computed with a second constant-Q filter bank. The scattering transform follows this latter approach.

3 Wavelet Scattering Transform

A scattering transform recovers the information lost by a mel-frequency averaging with a cascade of wavelet decompositions and modulus operators [31]. It is locally translation invariant and stable to time-warping deformation. Important properties of constant-Q filter banks are first reviewed in the framework of a wavelet transform, and the scattering transform is introduced in Section 3.2.

3.1 Analytic Wavelet Transform and Modulus

Constant-Q filter banks compute a wavelet transform. We review the properties of complex analytic wavelet transforms and their modulus, which are used to calculate mel-frequency spectral coefficients.

A wavelet $\psi(t)$ is a band-pass filter with $\hat{\psi}(0) = 0$. We consider analytic wavelets such that $\hat{\psi}(\omega) = 0$ for $\omega < 0$. As a result, $\psi(t)$ is a complex quadrature phase wavelet. For any $\lambda > 0$, a dilated wavelet of center frequency λ is written

$$\psi_\lambda(t) = \lambda \psi(\lambda t) \quad \text{and hence} \quad \hat{\psi}_\lambda(\omega) = \hat{\psi}\left(\frac{\omega}{\lambda}\right). \quad (12)$$

The center frequency of $\hat{\psi}$ is normalized to 1. In the following, we denote by Q the number of wavelets per octave, which means that $\lambda = 2^{j/Q}$ for $j \in \mathbb{Z}$. The bandwidth of $\hat{\psi}$ is of the order of Q^{-1} , to cover the whole frequency axis with these band-pass wavelet filters. The support of $\hat{\psi}_\lambda(\omega)$ is centered in λ with a frequency bandwidth λ/Q whereas the energy of $\psi_\lambda(t)$ is concentrated around 0 in an interval of size $2\pi Q/\lambda$. To guarantee that this interval is smaller than T , we define ψ_λ with (12) only for $\lambda \geq 2\pi Q/T$. For $\lambda < 2\pi Q/T$, the lower frequency interval $[0, 2\pi Q/T]$ is covered with about $Q - 1$ equally-spaced filters $\hat{\psi}_\lambda$ with constant frequency bandwidth $2\pi/T$. For simplicity, these lower-frequency filters are still called wavelets. We denote by Λ the grid of all wavelet center frequencies λ .

The wavelet transform of x computes a convolution of x with a low-pass filter ϕ of frequency bandwidth $2\pi/T$, and convolutions with all higher-frequency wavelets ψ_λ for $\lambda \in \Lambda$:

$$Wx = \left(x \star \phi(t), x \star \psi_\lambda(t) \right)_{t \in \mathbb{R}, \lambda \in \Lambda}. \quad (13)$$

This time index t is not critically sampled as in wavelet bases so this representation is highly redundant. The wavelet ψ and the low-pass filter ϕ are designed to build filters which cover the whole frequency axis, which means that

$$A(\omega) = |\hat{\phi}(\omega)|^2 + \frac{1}{2} \sum_{\lambda \in \Lambda} \left(|\hat{\psi}_\lambda(\omega)|^2 + |\hat{\psi}_\lambda(-\omega)|^2 \right) \quad (14)$$

satisfies, for all $\omega \in \mathbb{R}$:

$$1 - \alpha \leq A(\omega) \leq 1 \quad \text{with} \quad \alpha < 1. \quad (15)$$

This condition implies that the wavelet transform W is a stable and invertible operator. Multiplying (15) by $|\hat{x}(\omega)|^2$ and applying the Plancherel formula [30] gives

$$(1 - \alpha)\|x\|^2 \leq \|Wx\|^2 \leq \|x\|^2, \quad (16)$$

where $\|x\|^2 = \int |x(t)|^2 dt$ and the squared norm of Wx sums all squared coefficients:

$$\|Wx\|^2 = \int |x \star \phi(t)|^2 dt + \sum_{\lambda \in \Lambda} \int |x \star \psi_\lambda(t)|^2 dt.$$

The upper bound (16) means that W is a contractive operator and the lower bound implies that it has a stable inverse. One can also verify that the pseudo-inverse of W recovers x with the following formula

$$x(t) = (x \star \phi) \star \bar{\phi}(t) + \sum_{\lambda \in \Lambda} \text{Real}\left((x \star \psi_\lambda) \star \bar{\psi}_\lambda(t)\right), \quad (17)$$

with reconstruction filters defined by

$$\hat{\bar{\phi}}(\omega) = \frac{\hat{\phi}^*(\omega)}{A(\omega)} \quad \text{and} \quad \hat{\bar{\psi}}_\lambda(\omega) = \frac{\hat{\psi}_\lambda^*(\omega)}{A(\omega)}, \quad (18)$$

where z^* is the complex conjugate of $z \in \mathbb{C}$. If $\alpha = 0$ in (15) then W is unitary, $\bar{\phi}(t) = \phi(-t)$ and $\bar{\psi}_\lambda(t) = \psi_\lambda^*(-t)$.

Following (11), mel-frequency spectrograms can be approximated using a non-linear wavelet modulus operator which removes the complex phase of all wavelet coefficients:

$$\widetilde{W}x = \left(x \star \phi(t), |x \star \psi_\lambda(t)| \right)_{t \in \mathbb{R}, \lambda \in \Lambda}. \quad (19)$$

A signal cannot be reconstructed from the modulus of its Fourier transform but the situation is different for a wavelet transform which is highly redundant. Despite the loss of phase, for particular families of analytic wavelets, one can prove that \widetilde{W} is an invertible operator with a continuous inverse [45]. The modulus thus does not lose information.

This operator \widetilde{W} is also contractive. Indeed, the wavelet transform W is contractive and the complex modulus is contractive in the sense that $||a| - |b|| \leq |a - b|$ for any $(a, b) \in \mathbb{C}^2$ so

$$\|\widetilde{W}x - \widetilde{W}x'\|^2 \leq \|Wx - Wx'\|^2 \leq \|x - x'\|^2.$$

If W is a unitary operator then $\|\widetilde{W}x\| = \|Wx\| = \|x\|$ so \widetilde{W} preserves the signal norm.

One may define an analytic wavelet with an octave resolution Q as $\psi(t) = e^{it} \theta(t)$ and hence $\hat{\psi}(\omega) = \hat{\theta}(\omega - 1)$ where $\hat{\theta}$ is the transfer function of a low-pass filter whose bandwidth is of the order of Q^{-1} . If θ is a Gaussian then ψ is a complex Gabor function which is almost analytic because $|\hat{\psi}(\omega)|$ is small

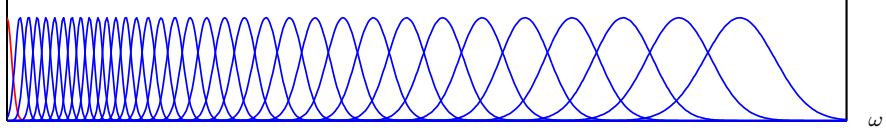


Figure 3: Gabor wavelets $\hat{\psi}_\lambda(\omega)$ with $Q = 8$ wavelets per octave, for different λ . The low frequency filter $\hat{\phi}(\omega)$ (in red) is a Gaussian.

but not strictly zero for $\omega < 0$. Figure 3 shows Gabor wavelets $\hat{\psi}_\lambda$ with $Q = 8$. In this case ϕ is also a Gaussian. Morlet wavelets are modified Gabor wavelets $\hat{\psi}(\omega) = \hat{\theta}(\omega - 1) - \hat{\theta}(\omega)\hat{\theta}(1)/\hat{\theta}(0)$ which guarantees that $\hat{\psi}(0) = 0$, and ϕ remains a Gaussian. For $Q = 1$, unitary wavelet transforms can also be obtained by choosing ψ to be the analytic part of a real wavelet which generates an orthogonal wavelet basis, such as a cubic spline wavelet [31].

3.2 Deep Scattering Network

We showed in (11) that mel-frequency spectral coefficients $Mx(t, \lambda)$ are approximately equal to averaged squared wavelet coefficients $|x \star \psi_\lambda|^2 \star |\phi|^2(t)$. The time averaging by the low-pass filter ϕ provides descriptors that are locally invariant to small translations relative to the duration T of ϕ . To avoid amplifying large amplitude coefficients, a scattering transform computes $|x \star \psi_\lambda| \star \phi(t)$ instead. It then recovers the information lost to time averaging by calculating additional invariant coefficients with a cascade of wavelet modulus transforms.

The simplest locally invariant descriptor of x is given by its time-average $S_0x(t) = x \star \phi(t)$, which removes all high frequencies. Complementary high-frequency information are provided by a first wavelet modulus transform

$$\widetilde{W}_1x = \left(x \star \phi(t), |x \star \psi_{\lambda_1}(t)| \right)_{t \in \mathbb{R}, \lambda_1 \in \Lambda_1},$$

computed with wavelets ψ_{λ_1} having an octave frequency resolution Q_1 . For audio signals we set $Q_1 = 8$, which defines wavelets having the same frequency resolution as mel-frequency filters. Audio signals have little energy at low frequencies so $S_0x(t) \approx 0$. Mel-frequency spectral coefficients are obtained by averaging the wavelet modulus coefficients with ϕ :

$$S_1x(t, \lambda_1) = |x \star \psi_{\lambda_1}| \star \phi(t). \quad (20)$$

These are called first-order scattering coefficients. They are computed with a second wavelet modulus transform applied to each $|x \star \psi_{\lambda_1}|$, which also provides complementary high frequency wavelet coefficients:

$$\widetilde{W}_2|x \star \psi_{\lambda_1}| = \left(|x \star \psi_{\lambda_1}| \star \phi, ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \right)_{\lambda_2 \in \Lambda_2}.$$

The wavelets ψ_{λ_2} have an octave resolution Q_2 which may be different from Q_1 . It is chosen to get a sparse representation which means concentrating the signal

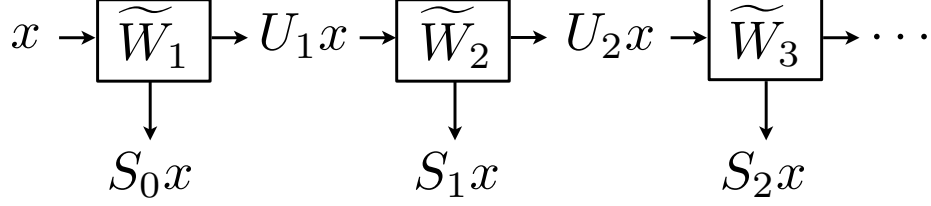


Figure 4: A scattering transform iterates on wavelet modulus operators \widetilde{W}_m to compute cascades of m wavelet convolutions and modulus stored in U_mx , and to output averaged scattering coefficients S_mx .

information over as few wavelet coefficients as possible. These coefficients are averaged by ϕ to obtain translation invariant coefficients, which defines second-order scattering coefficients:

$$S_2x(t, \lambda_1, \lambda_2) = ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) .$$

These averages are computed by applying a third wavelet modulus transform \widetilde{W}_3 to each $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$. It computes their wavelet coefficients through convolutions with a new set of wavelets ψ_{λ_3} having an octave resolution Q_3 . Iterating this process defines scattering coefficients at any order m .

For any $m \geq 1$, iterated wavelet modulus convolutions are written:

$$U_mx(t, \lambda_1, \dots, \lambda_m) = ||x \star \psi_{\lambda_1}| \star \dots| \star \psi_{\lambda_m}(t)| ,$$

where m th-order wavelets ψ_{λ_m} have an octave resolution Q_m , and satisfy the stability condition (15). Averaging U_mx with ϕ gives scattering coefficients of order m :

$$\begin{aligned} S_mx(t, \lambda_1, \dots, \lambda_m) &= ||x \star \psi_{\lambda_1}| \star \dots| \star \psi_{\lambda_m}| \star \phi(t) \\ &= U_mx(\cdot, \lambda_1, \dots, \lambda_m) \star \phi(t) . \end{aligned}$$

Applying \widetilde{W}_{m+1} on U_mx computes both S_mx and $U_{m+1}x$:

$$\widetilde{W}_{m+1}U_mx = (S_mx, U_{m+1}x) . \quad (21)$$

A scattering decomposition of maximal order \overline{m} is thus defined by initializing $U_0x = x$, and recursively computing (21) for $0 \leq m \leq \overline{m}$. This scattering transform is illustrated in Figure 4. The final scattering vector aggregates all scattering coefficients for $0 \leq m \leq \overline{m}$:

$$Sx = (S_mx)_{0 \leq m \leq \overline{m}} . \quad (22)$$

The scattering cascade of convolutions and non-linearities can also be interpreted as a convolutional network [25], where U_mx is the set of coefficients of the m th internal network layer. These networks have been shown to be highly

effective for audio classification [3, 14, 21, 24, 27, 34]. However, unlike standard convolutional networks, each such layer has an output $S_mx = U_mx \star \phi$, not just the last layer. In addition, all filters are predefined wavelets and are not learned from training data.

The wavelet octave resolutions are optimized at each layer m to produce sparse wavelet coefficients at the next layer. This better preserves the signal information as explained in Section 5. For audio signals x , we choose $Q_1 = 8$ wavelets per octave, which corresponds to a mel-frequency decomposition. This configuration has been shown to provide sparse representations of a mix of speech, music and environmental signals [41]. Such signals x often include harmonics which have sparse representations if the mother wavelet has sufficiently narrow frequency support. In addition, sparse representations have been shown to be better suited for classification [22, 35].

At the second order, choosing $Q_2 = 1$ defines wavelets with more narrow time support, which are better adapted to characterize transients and attacks in the second-order modulation spectrum. Section 6 shows that musical signals including modulation structures such as tremolo may however require wavelets having better frequency resolution, and hence $Q_2 > 1$. At higher orders $m \geq 3$ we set $Q_m = 1$ in all cases, but we will see that these coefficients can often be neglected.

The scattering cascade has similarities with several neurophysiological models of auditory processing, which incorporate cascades of constant-Q filter banks followed by non-linearities [11, 15]. The first filter bank with $Q_1 = 8$ models the cochlear filtering, whereas the second filter bank corresponds to later processing in the auditory pathway, which are modeled using filters with $Q_2 = 1$ [11, 15].

4 Scattering Properties

We briefly review important properties of scattering transforms, including stability to time-warping deformation, energy conservation and an algorithm for fast computation.

4.1 Time-Warping Stability

The Fourier transform is unstable to deformation because dilating a sinusoidal wave yields a new sinusoidal wave of different frequency which is orthogonal to the original one. Section 2 explains that mel-frequency spectrograms become stable to time-warping deformation with a frequency averaging. One can prove [31] that a scattering representation $\Phi(x) = Sx$ satisfies the Lipschitz continuity condition (3) relative to deformations because wavelet transforms are stable to deformation. Indeed, wavelets are regular and well-localized in time, and a small deformation of a wavelet yields a function which is highly similar to the original one.

The squared Euclidean norm of a scattering vector Sx is the sum of its

coefficients squared at all orders:

$$\begin{aligned}\|Sx\|^2 &= \sum_{m=0}^{\bar{m}} \|S_m x\|^2 \\ &= \sum_{m=0}^{\bar{m}} \sum_{\lambda_1, \dots, \lambda_m} \int |S_m x(t, \lambda_1, \dots, \lambda_m)|^2 dt.\end{aligned}$$

We consider deformations $x_\tau(t) = x(t - \tau(t))$ with $|\tau'(t)| < 1$ and $\sup_t |\tau(t)| \ll T$, which means that the maximum displacement is small relatively to the bandwidth of ϕ . One can prove [31] that there exists a constant C such that for all x and any such τ :

$$\|Sx_\tau - Sx\| \leq C \sup_t |\tau'(t)| \|x\|, \quad (23)$$

up to second-order terms. This Lipschitz continuity property implies that time-warping deformations can be locally linearized in the scattering space. Indeed, Lipschitz continuous operators are almost everywhere differentiable. Invariants to small deformations can thus be computed with linear operators in the scattering domain. This property is particularly important for linear discriminant classifiers such as support vector machines (SVMs).

4.2 Contraction and Energy Conservation

We show that a scattering transform is contractive and can preserve energy. Let us define the squared Euclidean norm $\|Ax\|^2$ of a vector of coefficients Ax (such as $W_m x$, $S_m x$, $U_m x$ or Sx) as the sum of all its coefficients squared.

Since Sx is computed by cascading wavelet modulus operators \widetilde{W}_m , which are all contractive, it results that S is also contractive:

$$\|Sx - Sx'\| \leq \|x - x'\|. \quad (24)$$

A scattering transform is therefore stable to additive noise.

If each wavelet transform is unitary, each \widetilde{W}_m preserves the signal norm. Applying this property to $\widetilde{W}_{m+1} U_m x = (S_m x, U_{m+1} x)$ yields

$$\|U_m x\|^2 = \|S_m x\|^2 + \|U_{m+1} x\|^2. \quad (25)$$

Summing these equations $0 \leq m \leq \bar{m}$ proves that

$$\|x\|^2 = \|Sx\|^2 + \|U_{\bar{m}+1} x\|^2. \quad (26)$$

Under appropriate assumptions on the mother wavelet ψ , one can prove that $\|U_{\bar{m}+1} x\|$ goes to zero as \bar{m} increases [31], which implies that $\|Sx\| = \|x\|$ for $\bar{m} = \infty$. This property comes from the fact that the modulus of analytic wavelet coefficients computes a smooth envelope, and hence pushes energy towards lower frequencies. By iterating on wavelet modulus operators, the scattering transform progressively propagates all the energy of $U_m x$ towards lower frequencies, which is captured by the low-pass filter of scattering coefficients $S_m x = U_m x \star \phi$.

T	$m = 0$	$m = 1$	$m = 2$	$m = 3$
23ms	0.1%	97.4%	2.4%	0.2%
93ms	0.0%	89.6%	8.3%	0.7%
370ms	0.0%	71.7%	23.7%	2.8%
1.5 s	0.0%	57.3%	32.9%	7.2%

Table 1: Averaged values $\|S_m x\|^2/\|x\|^2$ computed for signals x in the GTZAN music dataset [43], as a function of order m and averaging scale T . For $m = 1$, $S_m x$ is calculated by Gabor wavelets with $Q_1 = 8$, and for $m = 2, 3$ by cubic spline wavelets with $Q_2 = Q_3 = 1$.

One can verify numerically that $\|U_{\overline{m}+1}x\|$ converges to zero exponentially when \overline{m} goes to infinity and hence that $\|Sx\|$ converges exponentially to $\|x\|$. Table 1 gives the fraction of energy $\|S_m x\|^2/\|x\|^2$ absorbed by each scattering order. Since audio signals have little energy at low frequencies, $S_0 x$ is very small and most of the energy is absorbed by $S_1 x$ for $T = 23\text{ms}$. This explains why mel-frequency spectrograms are typically sufficient at these small time scales. However, as T increases, a progressively larger proportion of energy is absorbed by higher-order scattering coefficients. For $T = 370\text{ms}$, about 24% of the signal energy is captured in $S_2 x$. Section 6 shows that at this time scale, important amplitude modulation information is carried by these second-order coefficients. For $T = 370\text{ms}$, $S_3 x$ carries only 3% of the signal energy. It increases as T increases, but for audio classification applications studied in this paper, T remains below 370ms, so these third-order coefficients have a negligible role. We therefore concentrate on second-order scattering representations:

$$Sx = \left(S_0 x(t), S_1 x(t, \lambda_1), S_2 x(t, \lambda_1, \lambda_2) \right)_{t, \lambda_1, \lambda_2}. \quad (27)$$

4.3 Fast Scattering Computation

Subsampling scattering vectors provide a reduced representation, which leads to a faster implementation. Since the averaging window ϕ has a duration T , we compute scattering vectors at $t = kT/2$ for every integer k .

We suppose that $x(t)$ has N samples over each frame of duration T , and is thus sampled at a rate N/T . For each time frame $t = kT/2$, the number of first-order filters is about $Q_1 \log_2 N$ so there are about $Q_1 \log_2 N$ first-order coefficients $S_1 x(t, \lambda_1)$. We now show that the number of non-negligible second-order coefficients $S_2 x(t, \lambda_1, \lambda_2)$ is about $Q_1 Q_2 (\log_2 N)^2/2$.

The wavelet transform envelope $|x \star \psi_{\lambda_1}(t)|$ is a demodulated signal having approximatively the same frequency bandwidth as $\hat{\psi}_{\lambda_1}$. Its Fourier transform is mostly supported in the interval $[-\lambda_1 Q_1^{-1}, \lambda_1 Q_1^{-1}]$ for $\lambda_1 \geq 2\pi Q_1/T$, and in $[-2\pi T^{-1}, 2\pi T^{-1}]$ for $\lambda_1 \leq 2\pi Q_1/T$. If the support of $\hat{\psi}_{\lambda_2}$ centered at λ_2 does

not intersect the frequency support of $|x \star \psi_{\lambda_1}|$, then

$$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \approx 0 .$$

One can verify that non-negligible second-order coefficients satisfy

$$\lambda_2 \leq \max(\lambda_1 Q_1^{-1}, 2\pi T^{-1}) . \quad (28)$$

For a fixed t , a direct calculation then shows that there are of the order of $Q_1 Q_2 (\log_2 N)^2 / 2$ second-order scattering coefficients. Similar reasoning extends this result to show that there are about $Q_1 \dots Q_m (\log_2 N)^m / m!$ non-negligible m th-order scattering coefficients.

To compute $S_1 x$ and $S_2 x$ we first calculate $U_1 x$ and $U_2 x$ and average them with ϕ . Over a time frame of duration T , to reduce computations while avoiding aliasing, $|x \star \psi_{\lambda_1}(t)|$ is subsampled at a rate which is twice its bandwidth. The family of filters $\{\widehat{\psi}_{\lambda_1}\}_{\lambda_1 \in \Lambda_1}$ covers the whole frequency domain and Λ_1 is chosen so that filter supports barely overlap. Over a time frame where x has N samples, $\{|x \star \psi_{\lambda_1}(t)|\}_{\lambda_1 \in \Lambda_1}$ then has about $2N$ samples. Similarly, $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t)|$ is subsampled in time at a rate twice its bandwidth. The total number of samples for all λ_1 and λ_2 stays about $2N$. With an FFT, all first- and second-order wavelet modulus coefficients and their time averages $S_1 x(t, \lambda_1)$ and $S_2 x(t, \lambda_1, \lambda_2)$ are calculated for $t = kT/2$ with $O(N \log N)$ operations.

5 Inverse Scattering

To better understand the information carried by scattering coefficients, this section studies a numerical inversion of the transform. Since a scattering transform is computed by cascading wavelet modulus operators \widetilde{W}_m , the inversion approximatively inverts each \widetilde{W}_m for $m < \overline{m}$. At the maximum depth $m = \overline{m}$, the algorithm begins with a deconvolution, estimating $U_{\overline{m}} x(t)$ at all t on the sampling grid of $x(t)$, from $S_{\overline{m}} x(kT/2) = U_{\overline{m}} x \star \phi(kT/2)$.

Because of the subsampling, one cannot compute $U_{\overline{m}} x$ from $S_{\overline{m}} x$ exactly. This deconvolution is thus the main source of error. To take advantage of the fact that $U_{\overline{m}} x \geq 0$, the deconvolution is computed with the Richardson-Lucy algorithm [29], which preserves positivity if $\phi \geq 0$. We initialize $y_0(t)$ by interpolating $S_{\overline{m}} x(kT/2)$ linearly on the sampling grid of x , which introduces error because of aliasing. The Richardson-Lucy deconvolution iteratively computes

$$y_{n+1}(t) = y_n(t) \cdot \left[\left(\frac{y_0}{y_n \star \phi} \right) \star \tilde{\phi}(t) \right] , \quad (29)$$

with $\tilde{\phi}(t) = \phi(-t)$. It converges to the pseudo-inverse of the convolution operator applied to y_0 , which blows up when n increases because of the deconvolution instability. Deconvolution algorithms thus stop the iterations after a fixed number of iterations, which is set to 30 in this application.

Once an estimation of $U_{\overline{m}} x$ is calculated by deconvolution, we compute an estimate \tilde{x} of x by inverting each \widetilde{W}_m for $\overline{m} \geq m > 0$. As explained in Section

3.1, $\widetilde{W}x = (x \star \phi, |x \star \psi_\lambda|)_{\lambda \in \Lambda}$ can be inverted by taking advantage of the fact that wavelet coefficients define a complete and redundant signal representation. If $|x \star \psi_\lambda(t)| = 0$ then no phase needs to be recovered. The inversion of \widetilde{W} is thus more stable when $\widetilde{W}x$ is sparse, which motivates using wavelets ψ_{λ_m} providing a sparse representation at each order m . The inversion of \widetilde{W} amounts to solving a non-convex optimization problem. Recent convex relaxation approaches [7, 44] are able to compute exact solutions, but they require too much computation and memory for audio applications. Since the main source of errors is introduced at the deconvolution stage, one can use an approximate but fast inversion algorithm.

Griffin & Lim [19] showed that an alternating projection algorithm can recover good quality audio signals from their spectrograms, but with large mean-square errors because the algorithm is trapped in local minima. To compute an estimation \tilde{x} of x from $\widetilde{W}x$, this algorithm initializes \tilde{x}_0 to be a Gaussian white noise. For any $n \geq 0$, \tilde{x}_{n+1} is calculated from \tilde{x}_n by adjusting the modulus of its wavelet coefficients

$$z_\lambda(t) = |x \star \psi_\lambda(t)| \frac{\tilde{x}_n \star \psi_\lambda(t)}{|\tilde{x}_n \star \psi_\lambda(t)|} \quad (30)$$

and by applying the wavelet transform pseudo-inverse (17)

$$\tilde{x}_{n+1} = x \star \phi \star \bar{\phi}(t) + \sum_{\lambda \in \Lambda} \text{Real}\left(z_\lambda \star \bar{\psi}_\lambda(t)\right). \quad (31)$$

The dual filters are defined in (18). Numerical experiments are performed with $n = 32$ iterations, and we set $\tilde{x} = \tilde{x}_n$.

When $\overline{m} = 1$, an approximation \tilde{x} of x is computed from from (S_0x, S_1x) by first estimating U_1x from $S_1x = U_1x \star \phi$ with the Richardson-Lucy deconvolution algorithm. We then compute \tilde{x} from S_0x and this estimation of U_1x by approximatively inverting \widetilde{W}_1 with the Griffin & Lim algorithm. When T is above 100ms, the deconvolution loses too much information, and crude audio reconstructions are obtained from first-order coefficients. Figure 5(a) shows the scalograms $\log |x \star \psi_{\lambda_1}(t)|$ of a speech and a music signal, and the scalograms $\log |\tilde{x} \star \psi_{\lambda_1}(t)|$ of their approximations \tilde{x} from first-order scattering coefficients.

When $\overline{m} = 2$, the approximation \tilde{x} is calculated from (S_0x, S_1x, S_2x) by applying the deconvolution algorithm to $S_2x = U_2x \star \phi$ to estimate U_2x , and then by successively inverting \widetilde{W}_2 and \widetilde{W}_1 with the Griffin & Lim algorithm. Figure 5(c) shows $\log |\tilde{x} \star \psi_{\lambda_1}(t)|$ for the same speech and music signals. Amplitude modulations, vibratos and attacks are restored with greater precision by incorporating second-order coefficients, yielding a much better audio quality compared to first-order reconstructions. However, even with $\overline{m} = 2$, reconstructions become crude for $T \geq 500$ ms. Indeed, the number of second-order scattering coefficients $Q_1 Q_2 \log_2^2 N / 2$ is too small relatively to the number N audio samples in each audio frame, and they do not capture enough information. Examples of audio reconstructions are available at <http://www.di.ens.fr/signal/scattering/audio/>.

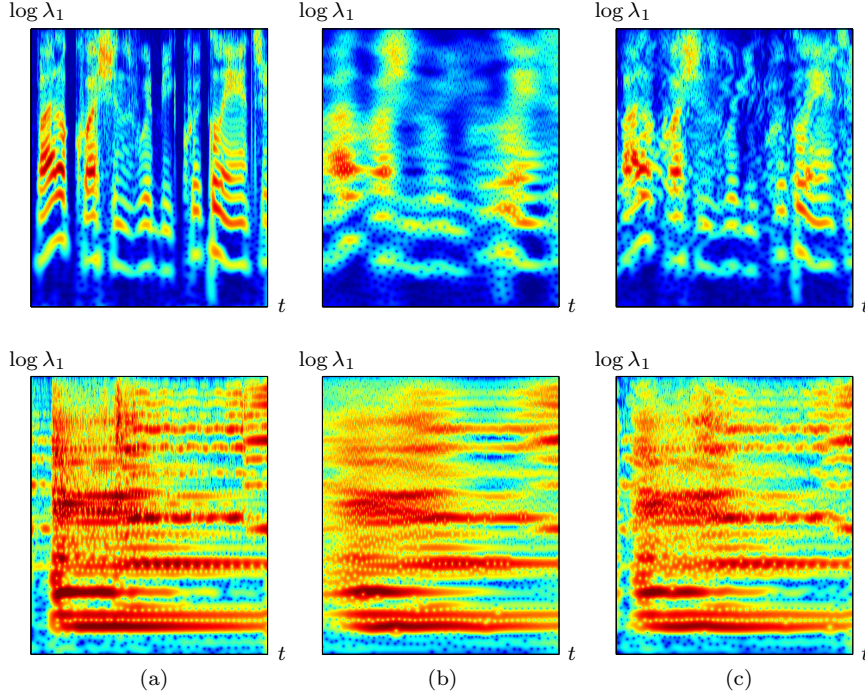


Figure 5: (a): Scalogram $\log |x \star \psi_{\lambda_1}(t)|$ for recordings of speech (top) and a cello (bottom). (b,c): Scalograms $\log |\tilde{x} \star \psi_{\lambda_1}(t)|$ of reconstructions \tilde{x} from first-order scattering coefficients ($\overline{m} = 1$) in (b), and from first- and second-order coefficients ($\overline{m} = 2$) in (c). Scattering coefficients were computed with $T = 190\text{ms}$ for the speech signal and $T = 370\text{ms}$ for the cello signal.

6 Scattering Spectrum

Whereas first-order scattering coefficients provide average spectrum measurements, which are equivalent to the mel-frequency spectrogram, second-order coefficients contain important complementary information. Section 6.1 shows that they provide better spectral resolution through interference measurements within each mel-scale interval. Section 6.2 proves that $S_2x(t, \lambda_1, \lambda_2)/S_1x(t, \lambda_1)$ characterizes the modulation spectrum of audio signals. As with MFCCs, computing the logarithm of scattering coefficients linearly separates all multiplicative components.

6.1 Frequency Interval Measurement from Interference

A wavelet transform has a worse frequency resolution than a windowed Fourier transform at high frequencies. However, we show that frequency intervals between harmonics are accurately measured by second-order scattering coefficients.

To simplify explanations, we consider a signal x of period T_0 . Squared wavelet coefficients can be written

$$|x \star \psi_{\lambda_1}(t)|^2 = \bar{e}^2 + \epsilon(t) , \quad (32)$$

where e^2 is the filtered signal energy

$$\bar{e}^2 = \frac{1}{T_0} \int_0^{T_0} |x \star \psi_{\lambda_1}(t)|^2 dt$$

and $\epsilon(t)$ is an oscillatory interference term giving the correlation between the frequency components in the support of $\hat{\psi}_{\lambda_1}$. If $\epsilon \ll \bar{e}^2$ then a first-order approximation of the square root applied to (32) gives

$$|x \star \psi_{\lambda_1}(t)| \approx \bar{e} + \frac{\epsilon(t)}{2\bar{e}} .$$

If we assume that the size T of the window ϕ satisfies $T \gg T_0$ then $\epsilon \star \phi \approx 0$ because ϵ contains only frequencies larger than $2\pi/T_0$, which are thus outside the frequency support $[-\pi/T, \pi/T]$ of $\hat{\phi}$. It results that

$$S_1 x(t, \lambda_1) = |x \star \psi_{\lambda_1}| \star \phi(t) \approx \bar{e} , \quad (33)$$

and $S_2 x(t, \lambda_1, \lambda_2) = |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2} \star \phi(t)$ satisfies

$$S_2 x(t, \lambda_1, \lambda_2) \approx \frac{1}{2\bar{e}} |\epsilon \star \psi_{\lambda_2}| \star \phi(t). \quad (34)$$

For example, if x has two frequency components in the support of $\hat{\psi}_{\lambda_1}$, we have

$$x \star \psi_{\lambda_1}(t) = \alpha_1 e^{i\xi_1 t} + \alpha_2 e^{i\xi_2 t}$$

and (32) then implies that $\epsilon(t) = 2\alpha_1 \alpha_2 \cos(\xi_1 - \xi_2)t$. Hence

$$\frac{S_2 x(t, \lambda_1, \lambda_2)}{S_1 x(t, \lambda_1)} \approx |\hat{\psi}_{\lambda_2}(\xi_2 - \xi_1)| \frac{\alpha_1 \alpha_2}{|\alpha_1|^2 + |\alpha_2|^2} .$$

These normalized second-order coefficients are thus non-negligible when λ_2 is of the order of the distance $|\xi_2 - \xi_1|$ between the two harmonics. This shows that although the first wavelet $\hat{\psi}_{\lambda_1}$ does not have enough resolution to discriminate the frequencies ξ_1 and ξ_2 , second-order coefficients detect their presence and accurately measure the interval $|\xi_2 - \xi_1|$. As in audio perception, scattering coefficients can accurately measure frequency intervals but not frequency location.

If $x \star \psi_{\lambda_1}(t) = \sum_n \alpha_n e^{i\xi_n t}$ has more frequency components, we verify similarly that $S_2 x(t, \lambda_1, \lambda_2)/S_1 x(t, \lambda_1)$ is non-negligible when λ_2 is of the order of $|\xi_n - \xi_{n'}|$ for some $n \neq n'$. These coefficients can thus measure multiple frequency intervals within the frequency band covered by $\hat{\psi}_{\lambda_1}$. If the frequency

resolution of $\widehat{\psi}_{\lambda_2}$ is not sufficient to discriminate between two frequency intervals $|\xi_1 - \xi_2|$ and $|\xi_3 - \xi_4|$, these intervals will interfere and create high amplitude third-order scattering coefficients. A similar calculation shows that third-order scattering coefficients $S_3x(t, \lambda_1, \lambda_2, \lambda_3)$ detect the presence of two such intervals within the support of $\widehat{\psi}_{\lambda_2}$ when λ_3 is close to $||\xi_1 - \xi_2| - |\xi_3 - \xi_4||$. They thus measure “intervals of intervals.”

Figure 6(a) shows the scalogram $\log|x \star \psi_{\lambda_1}|$ of a signal x containing a chord with two notes, whose fundamental frequencies are $\xi_1 = 600\text{Hz}$ and $\xi_2 = 675\text{Hz}$, followed by an arpeggio of the same two notes. First-order coefficients $\log S_1x(t, \lambda_1)$ in Figure 6(b) are very similar for the chord and the arpeggio because the time averaging loses time localization. However they are easily differentiated in Figure 6(c), which displays $\log(S_2x(t, \lambda_1, \lambda_2)/S_1x(t, \lambda_1))$ for $\lambda_1 \approx \xi_1 = 600\text{Hz}$, as a function of λ_2 . The chord creates large amplitude coefficients for $\lambda_2 = \xi_2 - \xi_1 = 75\text{Hz}$, which disappear for the arpeggio because these two frequencies are not present simultaneously. Second-order coefficients have also a large amplitude at low frequencies λ_2 . These arise from variation of the note envelopes in the chord and in the arpeggio, as explained in the next section.

6.2 Amplitude Modulation Spectrum

Audio signals are usually modulated in amplitude by an envelope, whose variations may correspond to an attack or a tremolo. For voiced and unvoiced sounds modeled by harmonic sounds and Gaussian noises, we show that amplitude modulations are characterized by second-order scattering coefficients.

Let $x(t)$ be a sound resulting from an excitation $e(t)$ filtered by a resonance cavity of impulse response $h(t)$, which is modulated in amplitude by $a(t) \geq 0$ to give

$$x(t) = a(t) (e \star h)(t) . \quad (35)$$

The impulse response h is typically very short compared to the minimum variation interval $(\sup_t |a'(t)|)^{-1}$ of the modulation term. Observe that if $\lambda_1 \geq 2\pi Q_1/T$ satisfies

$$\left(\int |t| |h(t)| dt \right)^{-1} \gg \frac{\lambda_1}{Q_1} \gg \sup_t |a'(t)| , \quad (36)$$

then $a(t)$ remains nearly constant over the time support of ψ_{λ_1} and $\widehat{h}(\omega)$ is nearly constant over the frequency support of $\widehat{\psi}_{\lambda_1}$. It results that

$$|x \star \psi_{\lambda_1}(t)| \approx |\widehat{h}(\lambda_1)| |e \star \psi_{\lambda_1}(t)| a(t) . \quad (37)$$

We shall compute $|e \star \psi_{\lambda_1}|$ when $e(t)$ is a pulse train or a Gaussian white noise and derive the values of first- and second-order scattering coefficients.

For a voiced sound, the excitation is modeled by a pulse train of pitch ξ :

$$e(t) = \frac{\xi}{2\pi} \sum_n \delta \left(t - \frac{2n\pi}{\xi} \right) = \sum_k e^{ik\xi t} .$$

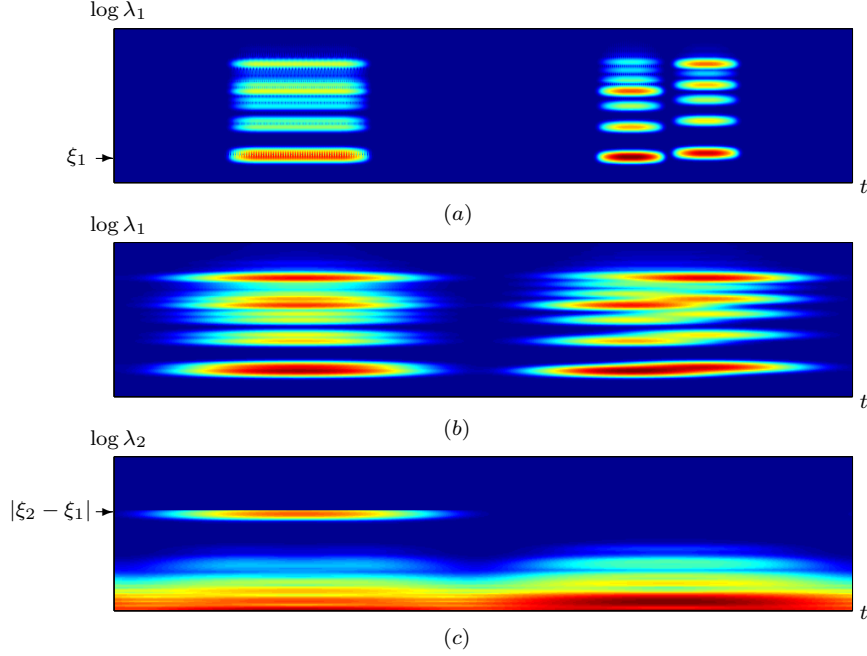


Figure 6: (a): Scalogram $\log |x \star \psi_{\lambda_1}(t)|$ for a signal with two notes, of fundamental frequencies $\xi_1 = 600\text{Hz}$ and $\xi_2 = 675\text{Hz}$, first played as a chord and then as an arpeggio. (b): First-order scattering coefficients $\log S_1 x(t, \lambda_1)$ for $T = 512\text{ms}$. (c): Second-order scattering coefficients $\log(S_2(t, \xi_1, \lambda_2)/S_1(t, \lambda_1))$ with $\lambda_1 = \xi_1$ as a function of t and λ_2 . The chord interferences produce large coefficients for $\lambda_2 = |\xi_2 - \xi_1|$.

Suppose that $\lambda_1/Q_1 \ll \xi$ so that the support of $\widehat{\psi}_{\lambda_1}$ covers at most one partial, whose frequency $k\xi$ is the closest to λ_1 . It then results from (37) that

$$|x \star \psi_{\lambda_1}(t)| \approx |\widehat{h}(\lambda_1)| |\widehat{\psi}_{\lambda_1}(k\xi)| a(t), \quad (38)$$

so $S_1 x(t, \lambda_1) = |x \star \psi_{\lambda_1}| \star \phi(t)$ is given by

$$S_1 x(t, \lambda_1) \approx |\widehat{h}(\lambda_1)| |\widehat{\psi}_{\lambda_1}(k\xi)| a \star \phi(t). \quad (39)$$

It is non-zero when λ_1 is close to a harmonic $k\xi$, and is proportional to $|\widehat{h}(\lambda_1)|$.

Figure 7(a) displays $\log |x \star \psi_{\lambda_1}(t)|$ for a signal having three voiced and three unvoiced sounds. The first three are produced by a pulse train excitation $e(t)$ with a pitch of $\xi = 600\text{ Hz}$. Figure 7(b) shows that $\log S_1 x(t, \lambda_1)$ has a harmonic structure, with an amplitude depending on $\log |\widehat{h}(\lambda_1)|$. However, the averaging by ϕ removes the differences between the modulation amplitudes $a(t)$ of these three voiced sounds.

Using (38), we compute second-order scattering coefficients

$$S_2 x(t, \lambda_1, \lambda_2) \approx |\widehat{h}(\lambda_1)| |\widehat{\psi}_{\lambda_1}(k\xi)| |a \star \psi_{\lambda_2}| \star \phi(t),$$

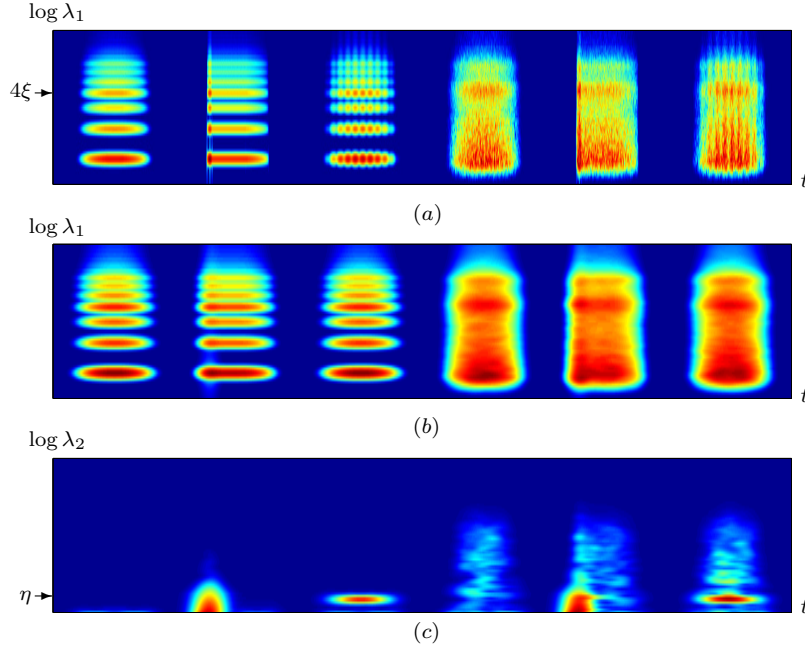


Figure 7: (a): $\log |x \star \psi_{\lambda_1}(t)|$ for a signal with three voiced sounds of same pitch $\xi = 600\text{Hz}$ and same $h(t)$ but different amplitude modulations $a(t)$: first a smooth attack, then a sharp attack, then a tremolo of frequency η . It is followed by three unvoiced sounds created with the same $h(t)$ and same amplitude modulations $a(t)$ as the first three voiced sounds. (b): First-order scattering $\log S_1x(t, \lambda_1)$ with $T = 128\text{ms}$. (c): Second-order scattering $\log(S_2x(t, \lambda_1, \lambda_2)/S_1x(t, \lambda_1))$ displayed for $\lambda_1 = 4\xi$, as a function of t and λ_2 .

which implies that

$$\frac{S_2x(t, \lambda_1, \lambda_2)}{S_1x(t, \lambda_1)} \approx \frac{|a \star \psi_{\lambda_2}| \star \phi(t)}{a \star \phi(t)}. \quad (40)$$

Normalized second-order scattering coefficients thus depend only on the amplitude modulation $a(t)$, and compute its wavelet spectrum at all frequencies λ_2 .

Figure 7(c) displays $\log(S_2(t, \lambda_1, \lambda_2)/S_1(t, \lambda_1))$ for the fourth partial $\lambda_1 = 4\xi$, as a function of λ_2 . The modulation envelope $a(t)$ of the first sound has a smooth attack and thus produces large coefficients only at low frequencies λ_2 . The envelope $a(t)$ of the second sound has a much sharper attack and thus produces large amplitude coefficients for higher frequencies λ_2 . The third sound is modulated by a tremolo, which is a periodic oscillation $a(t) = 1 + \epsilon \cos(\eta t)$. According to (40), this tremolo creates large amplitude coefficients when $\lambda_2 = \eta$, as shown in Figure 7(c).

Unvoiced sounds are modeled by excitations $e(t)$ which are realizations of Gaussian white noise. The modulation amplitude is typically non-sparse, which

means the square of the average of $a(t)$ on intervals of size T is of the order of the average of $a^2(t)$. If $\lambda_1 Q_1^{-1} \gg T^{-1}$ then Appendix A shows

$$S_1 x(t, \lambda_1) \approx \frac{\pi^{1/2}}{2} \|\psi\| \lambda_1^{1/2} |\hat{h}(\lambda_1)| a \star \phi(t) . \quad (41)$$

First-order scattering coefficients are again proportional to $|\hat{h}(\lambda_1)|$ but do not have a harmonic structure. This is shown in Figure 7(b) by the last three unvoiced sounds. The fourth, fifth, and sixth sounds have the same filter $h(t)$ and envelope $a(t)$ as the first, second, and third sounds, respectively, but with a Gaussian white noise excitation.

Appendix A also shows that if $a(t)$ is non-sparse and $\lambda_1 Q_1^{-1} \gg T^{-1}$ then

$$\frac{S_2 x(t, \lambda_1, \lambda_2)}{S_1 x(t, \lambda_1)} = \frac{|a \star \psi_{\lambda_2}| \star \phi(t)}{a \star \phi(t)} + \tilde{\epsilon}(t)$$

where $\tilde{\epsilon}(t)$ is small relatively to the first amplitude modulation term if $(4/\pi - 1)^{1/2} (\lambda_2 Q_1)^{1/2} (\lambda_1 Q_2)^{-1/2}$ is small relatively to this modulation term. Voiced and unvoiced sounds thus produce similar second-order scattering coefficients. This is illustrated by Figure 7(c), which shows that the fourth, fifth, and sixth sounds have second-order coefficients similar to those of the first, second, and third sounds, respectively. The stochastic error term $\tilde{\epsilon}$ produced by unvoiced sounds appears as small amplitude random fluctuations in Figure 7(c).

7 Frequency Transposition Invariance

Audio signals within the same class may be transposed in frequency, as when the same word is pronounced by a man or a woman. This frequency transposition is a complex phenomenon which affects the pitch and filter differently. The pitch is typically translated on a logarithmic frequency scale whereas filters are not just translated but also deformed. We thus need a representation which is invariant to frequency translation on a logarithmic scale, but also stable to frequency deformation. After reviewing the Mel-frequency cepstral coefficient (MFCC) approach through the discrete cosine transform (DCT), this section defines such a representation with a scattering transform computed along frequency.

MFCCs are computed from the mel-frequency spectrogram $\log Mx(t, \lambda)$ by calculating a DCT along the mel-frequency index γ for a fixed t [16]. This γ is linear in λ for low frequencies, but is proportional to $\log_2 \lambda$ for higher frequencies. For simplicity, we write $\gamma = \log_2 \lambda$ and $\lambda = 2^\gamma$, although this should be modified at low frequencies.

The frequency index of the DCT is called the “quefrequency” parameter. Setting high-quefrequency coefficients to zero is equivalent to averaging $\log Mx(t, 2^\gamma)$ along γ , which provides some frequency transposition invariance. The more high-quefrequency coefficients are set to zero, the bigger the averaging and hence the more transposition invariance obtained, but at the expense of losing information.

We avoid this loss by replacing the DCT with a scattering transform along γ . A frequency scattering transform is calculated by iteratively applying wavelet

transforms and modulus operators. An analytic wavelet transform of a log-frequency dependent signal $z(\gamma)$ is defined as in (13), but with convolutions along the log-frequency variable γ instead of time:

$$W^{\text{fr}} z = \left(z \star \phi^{\text{fr}}(\gamma), z \star \psi_q(\gamma) \right)_{\gamma, q}. \quad (42)$$

Each wavelet ψ_q is a band-pass filter whose Fourier transform $\widehat{\psi}_q$ is centered at “quefrequency” q and ϕ^{fr} is an averaging filter. These wavelets satisfy the condition (15), so W^{fr} is contractive and invertible.

A scattering transform computes a cascade of wavelet modulus transforms. Similarly to (27), we iteratively compute wavelet modulus convolutions

$$U^{\text{fr}} z = \left(z(\gamma), |z \star \psi_{q_1}(\gamma)|, ||z \star \psi_{q_1} \star \psi_{q_2}(\gamma)|| \right). \quad (43)$$

Averaging these defines a second-order scattering transform:

$$S^{\text{fr}} z = \left(z \star \phi^{\text{fr}}(\gamma), |z \star \psi_{q_1} \star \phi^{\text{fr}}(\gamma)|, ||z \star \psi_{q_1} \star \psi_{q_2} \star \phi^{\text{fr}}(\gamma)|| \right). \quad (44)$$

These coefficients are locally invariant to log-frequency shifts, over a domain proportional to the support of the averaging filter ϕ^{fr} . This frequency scattering is formally identical to a time scattering transform, having the same properties when exchanging time and log-frequency variables. Numerical experiments are implemented with Morlet wavelets ψ_{q_1} and ψ_{q_2} with $Q_1 = Q_2 = 1$.

The frequency scattering transform is not applied to MFCCs but to the logarithm of normalized first- and second-order time scattering coefficients, as computed in Section 3. For audio signals, $S_0 x = x \star \phi \approx 0$ so we neglect these coefficients. We also normalize the scattering transform by dividing second-order coefficients by their corresponding first-order coefficients. Abusing notation slightly, we still refer to this normalized scattering transform as Sx . As explained in Section 6, taking the logarithm separates signal components, such as amplitude modulations. As with MFCCs, the frequency scattering is computed along the log-frequency parameter γ . For a fixed frequency $\lambda_1 = 2^\gamma$ and time t , the second-order log-normalized time scattering vector is

$$\log Sx(t, \gamma) = \left(\begin{array}{c} \log(|x \star \psi_{2^\gamma}| \star \phi(t)) \\ \log\left(\frac{|x \star \psi_{2^\gamma}| \star \psi_{\lambda_2} \star \phi(t)}{|x \star \psi_{2^\gamma}| \star \phi(t)}\right) \end{array} \right)_{\lambda_2} \quad (45)$$

This is a multidimensional vector of signals $z(\gamma)$, depending only on γ for a fixed t and λ_2 . Let us transform each $z(\gamma)$ by the frequency scattering operators U^{fr} and S^{fr} defined in (43) and (44). We let $U^{\text{fr}} \log Sx(t, \gamma)$ and $S^{\text{fr}} \log Sx(t, \gamma)$ stand for the concatenation of these transformed signals for all t and λ_2 . The representation $S^{\text{fr}} \log Sx$ thus cascades a scattering in time and in frequency, so it is locally translation invariant in time and in log-frequency, as well as stable to time and frequency deformations. The interval of time invariance is defined by the size of the time averaging window ϕ , whereas its frequency invariance depends upon the width of the frequency averaging window ϕ^{fr} .

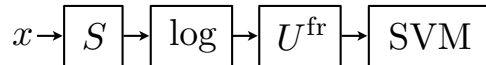


Figure 8: A time and frequency scattering representation is computed by applying a normalized temporal scattering S on the input signal $x(t)$, a logarithm, and a scattering along log-frequency without averaging.

For different tasks, frequency transposition may help classification or may destroy important information, the latter being the case for speaker identification. The size of the frequency averaging filter ϕ^{fr} should thus depend upon the task. Next section explains how this is learned at the classification stage.

8 Classification

This section compares the classification performance of support vector machine classifiers applied to scattering representations with standard low-level features such as Δ -MFCCs or state-of-the-art representations. Section 8.1 explains how to automatically adapt invariance parameters, while Sections 8.2 and 8.3 present results for musical genre classification and phone identification, respectively.

8.1 Adapting Time and Frequency Transposition Invariance

The amount of time shift and frequency transposition invariance depends on the classification problem, and may vary for each signal class. This adaptation is implemented by a supervised classifier, applied to the time and frequency scattering representation.

Figure 8 illustrates the computation of a time and frequency scattering representation. The scattering transform Sx of an input signal x is computed along time, with averaging scale T , and sampled at time intervals $T/2$. The transform is normalized and a logarithm is applied to separate multiplicative factors. Scattering coefficients are indexed by their log-frequency parameter γ , which defines the vector $\log Sx(t, \gamma)$ in (45). For a fixed t , $U^{\text{fr}} \log Sx(t, \gamma)$ is calculated as in (43), with wavelet convolutions along the parameter γ . Averaging $U^{\text{fr}} \log Sx(t, \gamma)$ with $\phi^{\text{fr}}(\gamma)$ computes a frequency scattering transform $S^{\text{fr}} \log Sx(t, \gamma)$, which is locally invariant to frequency transposition.

Since we do not know in advance how much transposition invariance is needed for a particular classification task, the final frequency averaging is adaptively computed by the supervised classifier, which takes for input $\{U^{\text{fr}} \log Sx(t, \gamma)\}_{\gamma}$ for each time frame t . The supervised classification is implemented by a support vector machine (SVM). A binary SVM classifies a feature vector by calculating its position relative to a hyperplane, which is optimized to maximize class separation given a set of training samples. It thus computes the sign of an optimized linear combination of the feature vector coefficients. With a Gaussian

kernel of variance σ^2 , the SVM computes different hyperplanes in different balls of radius σ in the feature space. The coefficients of the linear combination thus vary smoothly with the feature vector values. Applied to $\{U^{\text{fr}} \log Sx(t, \gamma)\}_\gamma$, the SVM optimizes the linear combination of coefficients along γ , and can thus adjust the amount of linear averaging to create frequency transposition invariant descriptors which maximize class separation. A multi-class SVM is computed from binary classifiers using a one-versus-one approach. All numerical experiments use the LIBSVM library [8].

We can also use this approach to automatically adjust the wavelet octave resolution Q_1 . We compute the time scattering for several values of Q_1 , and concatenate the coefficients in a single feature vector. A filter bank with $Q_1 = 8$ has enough frequency resolution to separate harmonic structures, whereas wavelets with $Q_1 = 1$ have a smaller time support and can thus better localize transient in time. By calculating linear combinations of feature vector coefficients, the SVM can amplify the coefficients corresponding to a better Q_1 , depending upon the type of structure needed to best discriminate a given class. In the experiments described below, adding more values of Q_1 between 1 and 8 provides marginal improvements.

Classification results can also be improved by adapting the averaging size T of the time scattering, for each signal class. For example, a phone duration may range from 10ms to 200ms and shorter phones are better discriminated with scattering coefficients calculated with smaller T . We thus concatenate scattering transforms computed for several T , letting the SVM amplify scattering coefficients computed with a T that is best adapted to each class. In the experiments, this adaptivity is implemented with three values of T .

8.2 Musical Genre Classification

Scattering feature vectors are first applied to musical genre classification problem on the GTZAN dataset [43]. The dataset consists of 1000 thirty-second clips, divided into 10 genres of 100 clips each. Given a clip, the goal is to find its genre.

A set of feature vectors is computed over half-overlapping audio frames of duration T . Each frame of a clip is classified separately by a Gaussian kernel SVM, and the clip is assigned to the class which is most often selected by its frames. To reduce the SVM training time, feature vectors were only computed every 370ms for the training set. The SVM slack parameter and the Gaussian kernel variance are determined through cross-validation on the training data. Table 2 summarizes results one run of ten-fold cross-validation. It gives the average error and its standard deviation.

A Δ -MFCC vector represents an audio frame of duration T at time t by three MFCC vectors centered at $t - T/2$, t and $t + T/2$. When computed for $T = 23\text{ms}$, the Δ -MFCC error is 19.3%, which is reduced to 17.8% by increasing T to 370ms. This is because the frames must be sufficiently long to include enough musical structure. Further increasing T does not reduce the error because the underlying ergodic stationarity hypothesis is strongly violated.

Representations	GTZAN	TIMIT
Δ -MFCC ($T = 23\text{ms}$)	19.3 ± 4.2	19.3
Δ -MFCC ($T = 370\text{ms}$)	17.8 ± 4.2	66.1
State of the art (excluding scattering)	9.4 ± 3.1 [26]	16.7 [9]
	$T = 370\text{ms}$	$T = 32\text{ms}$
Time Scat., $\overline{m} = 1$	17.9 ± 4.2	18.5
Time Scat., $\overline{m} = 2$	12.3 ± 2.7	17.7
Time Scat., $\overline{m} = 3$	10.7 ± 2.0	18.7
Time & Freq. Scat., $\overline{m} = 2$	10.3 ± 2.3	16.5
Adapt Q_1 , Time & Freq. Scat., $\overline{m} = 2$	9.0 ± 2.0	16.1
Adapt Q_1, T , Time & Freq. Scat., $\overline{m} = 2$	8.1 ± 2.3	15.8

Table 2: Error rates (in percent) for musical genre classification on GTZAN and for phone identification on the TIMIT database for different features. Time scattering transforms are computed with $T = 370\text{ms}$ for GTZAN and with $T = 32\text{ms}$ for TIMIT.

State-of-the-art algorithms provide refined feature vectors to improve classification. For example, combining MFCCs with stabilized modulation spectra and performing linear discriminant analysis, [26] obtains an error of 9.4%, the best result so far. In [21], a deep belief network is trained on spectrograms, achieving a 15.7% error with an SVM classifier. Finally, [22], sparse representation on a constant-Q transform, giving a 16.6% error using an SVM. A wavelet scattering does not need any learning because the nature of time and frequency invariants is known and leads to an optimized choice of wavelet filters. It reduces computations and improves results relatively to a learning approach.

Table 2 gives classification errors for different scattering feature vectors. For $\overline{m} = 1$, they are composed of first-order time scattering coefficients computed using Gabor wavelets with $Q_1 = 8$ and $T = 370\text{ms}$. These vectors are similar to an MFCCs as shown by (11). As a result, the classification error of 17.9% is close to that of MFCCs for the same T . For $\overline{m} = 2$, we add second-order coefficients computed using Morlet wavelets with $Q_2 = 2$. It reduces the error to 12.3%. This 30% error reduction shows the importance of second-order coefficients for relatively large T . Third-order coefficients are also computed with Morlet wavelets with $Q_3 = 1$. For $\overline{m} = 3$, including these coefficients reduces the error marginally to 10.7%, at a significant computational and memory cost. We thus restrict ourselves to $\overline{m} = 2$.

Musical genre recognition is a task which is partly invariant to frequency transposition. Incorporating a scattering along the log-frequency variable, for frequency transposition invariance, reduces the error by about 20%. These errors are obtained with a first-order scattering along log-frequency. Adding second-order coefficients improves results marginally.

Providing adaptivity for the wavelet octave bandwidth Q_1 by computing scattering coefficients for both $Q_1 = 1$ and $Q_1 = 8$ further reduces the error

by about 10%. Indeed, music signals include both sharp transients and narrow-bandwidth frequency components. Further enriching the representation by concatenating scattering coefficients for $T = 370\text{ms}$, 740ms , 1.5s also reduces the error rate, which is to be expected since musical signals contain structures at both short and long scales. This yields an error rate of 8.1%, which compares favorably to the non-scattering state-of-the-art of 9.4% error [26].

Replacing the SVM with more sophisticated classifiers can improve results. The error rate from second-order time-scattering coefficients is reduced from 12.3% to 8.8% in [10], with a sparse representation classifier.

8.3 Phone Recognition

The same scattering representation is tested for phone recognition with the TIMIT corpus [18]. The dataset contains 6300 phrases, each annotated with the identities, locations, and durations of its constituent phones. Given the location and duration of a phone, the goal is to determine its class according to the standard protocol [13, 28]. The 61 phone classes (excluding the glottal stop /q/) are collapsed into 48 classes, which are used to train and test models. To calculate the error rate, these classes are then mapped into 39 clusters. Training is achieved on the full 3696-phrase training set, excluding “SA” sentences. The Gaussian kernel SVM parameters are optimized by validation on the standard 400-phrase development set [20]. The error is then calculated on the core 192-phrase test set.

An audio segment of length 192ms centered on a phone can be represented as an array of MFCC feature vectors with half-overlapping time windows of duration T . This array, with the logarithm of the phone duration added, is fed to the SVM. Table 2 shows that $T = 23\text{ms}$ yields a 19.3% error which is much less than the 66.1% error for $T = 370\text{ms}$, since many phones have a short duration with highly transient structures.

A lower error of 17.1% is obtained by replacing the SVM with a sparse representation classifier on MFCC-like spectral features [38]. Combining MFCCs of different window sizes and using a committee-based hierarchical discriminative classifier, [9] achieves an error of 16.7%, the best so far. Finally, convolutional deep-belief networks cascades convolutions, similarly to scattering, on a spectrogram using filters learned from the training data. These, combined with MFCCs, yield an error of 19.7%.

Rows 4 through 6 of Table 2 gives the classification results obtained by replacing MFCC vectors with a time scattering transform computed with first-order Gabor wavelets with $Q_1 = 8$. Second- and third-order scattering coefficients are calculated with Morlet wavelets with $Q_2 = Q_3 = 1$. The best results are obtained with $T = 32\text{ms}$. For $\overline{m} = 1$, we only keep first-order scattering coefficients and get a 18.5% error, similar to that of MFCCs. The error is reduced by about 5% with $\overline{m} = 2$, a smaller improvement than for GTZAN because scattering invariants are computed on smaller time interval $T = 32\text{ms}$ as opposed to 370ms for music. Second-order coefficients carry less energy when T is smaller, as shown in Table 1. For the same reason, third-order coefficients provide even

less information compared to the GTZAN case, and do not improve results.

For $\overline{m} = 2$, cascading a log-frequency transposition invariance computed with a first-order frequency scattering transform of Section 7 reduces the error by about 5%. Computing a second-order frequency scattering transform only marginally improves results. Allowing to adapt the wavelet frequency resolution by computing scattering coefficients with $Q_1 = 1$ and $Q_1 = 8$ also reduces the error by a small amount. Finally, adapting the interval T further improves results because different phones often have very different durations. This is done by aggregating scattering coefficients computed for $T = 32\text{ms}$, 64ms , 128ms .

9 Conclusion

The success of MFCCs for audio classification can partially be explained by their stability to time-warping deformation. Scattering representations extend MFCCs by recovering lost high frequencies through successive wavelet convolutions. It provides modulation spectrum measurements which are stable to time-warping deformation, and it carries the whole signal energy. The logarithm of second-order scattering coefficients characterizes amplitude modulations, including transient phenomena such as attacks. Over $T \approx 200\text{ms}$, good audio signal quality is recovered from first- and second-order scattering coefficients. A frequency transposition invariant representation is obtained by cascading a second scattering transform along frequencies. Time and frequency scattering feature vectors yield state-of-the-art classification results with a Gaussian kernel SVM, for musical genre classification on GTZAN, and phone identification on TIMIT.

A Modulation Spectrum Properties

This appendix gives approximations of first- and second-order scattering coefficients produced by $x(t) = a(t) (e \star h)(t)$, for a Gaussian white noise excitation $e(t)$.

We saw in (37) that

$$|x \star \psi_{\lambda_1}(t)| \approx |\widehat{h}(\lambda_1)| |e \star \psi_{\lambda_1}(t)| a(t) . \quad (46)$$

Let us decompose

$$|e \star \psi_{\lambda_1}(t)| = \mathbb{E}(|e \star \psi_{\lambda_1}|) + \epsilon(t) , \quad (47)$$

where $\epsilon(t)$ is a zero-mean stationary process. Since $e(t)$ is a normalized Gaussian white noise, $e \star \psi_{\lambda_1}(t)$ is a Gaussian random variable of variance $\|\psi_{\lambda_1}\|^2$. It results that $|e \star \psi_{\lambda_1}(t)|$ and $\epsilon(t)$ have a Rayleigh distribution, and since ψ is a complex quadrature phase wavelet, one can verify that

$$\mathbb{E}(|e \star \psi_{\lambda_1}|)^2 = \frac{\pi}{4} \mathbb{E}(|e \star \psi_{\lambda_1}|^2) = \frac{\pi}{4} \|\psi_{\lambda_1}\|^2 .$$

Inserting (47) and this equation in (46) shows that

$$|x \star \psi_{\lambda_1}(t)| \approx |\widehat{h}(\lambda_1)| \left(\pi^{1/2} 2^{-1} \|\psi_{\lambda_1}\| a(t) + a(t) \epsilon(t) \right). \quad (48)$$

When averaging with ϕ , we get

$$S_1 x(t, \lambda_1) \approx |\widehat{h}(\lambda_1)| \left(\pi^{1/2} 2^{-1} \|\psi_{\lambda_1}\| a \star \phi(t) + (a \epsilon) \star \phi(t) \right). \quad (49)$$

We are going to show that if $T^{-1} \ll \lambda_1 Q_1^{-1}$ and $a(t)$ is not sparse, which means the square of the average of $a(t)$ on intervals of size T is of the order of the average of $a^2(t)$, then

$$\frac{\mathbb{E}(|(a \epsilon) \star \phi(t)|^2)}{\|\psi_{\lambda_1}\|^2 |a \star \phi(t)|^2} \ll 1 \quad (50)$$

which implies (41). We give the main arguments to compute the order of magnitudes of the stochastic terms, but it is not a rigorous proof. Computations rely on the following lemma.

Lemma 1. *Let $z(t)$ be a zero-mean stationary process of power spectrum $\widehat{R}_z(\omega)$. For any deterministic functions $a(t)$ and $h(t)$*

$$\mathbb{E}(|(za) \star h(t)|^2) \leq \sup_{\omega} \widehat{R}_z(\omega) |a|^2 \star |h|^2(t). \quad (51)$$

Proof. Let $R_z(\tau) = \mathbb{E}(z(t) z(t + \tau))$,

$$\mathbb{E}(|(za) \star h(t)|^2) = \iint R_z(v - u) a(u) h(t - u) a(v)^* h(t - v)^* du dv$$

and hence

$$\mathbb{E}(|(za) \star h(t)|^2) = \langle R_z y_t, y_t \rangle \text{ with } y_t(u) = a(u) h(t - u).$$

Since R_z is the kernel of a positive symmetric operator whose spectrum is bounded by $\sup_{\omega} \widehat{R}_z(\omega)$ it results that

$$\mathbb{E}(|(za) \star h(t)|^2) \leq \sup_{\omega} \widehat{R}_z(\omega) \|y_t\|^2 = \sup_{\omega} \widehat{R}_z(\omega) |a|^2 \star |h|^2(t).$$

□

Since $e(t)$ is a normalized white noise, $e \star \psi_{\lambda_1}$ is a Gaussian process and $\epsilon(t)$ is a stationary Rayleigh process. With a Gaussian chaos expansion, one can verify [6] that $\sup_{\omega} |\widehat{R}_{\epsilon}(\omega)| \leq 1 - \pi/4$. Applying Lemma 1 to $z = \epsilon$ and $h = \phi$ gives

$$\mathbb{E}(|(\epsilon a) \star \phi(t)|^2) \leq (1 - \pi/4) |a|^2 \star |\phi|^2(t).$$

Since ϕ has a duration T , it can be written as $\phi(t) = T^{-1} \phi_0(T^{-1}t)$ for some ϕ_0 of duration 1. As a result, if the square of the average of $a(t)$ is of the order of the average of $a^2(t)$ then

$$\frac{|a|^2 \star |\phi|^2(t)}{|a \star \phi(t)|^2} \sim \frac{1}{T} \quad (52)$$

The frequency support of ψ_{λ_1} is proportional to $\lambda_1 Q_1^{-1}$, so we have $\|\psi_{\lambda_1}\|^2 \sim \lambda_1 Q_1^{-1}$. Together with (52), if $T^{-1} \ll \lambda_1 Q_1^{-1}$ it proves (50) and hence

$$S_1 x(t, \lambda_1) \approx \frac{\pi^{1/2}}{2} \|\psi\| \lambda_1^{1/2} |\widehat{h}(\lambda_1)| a \star \phi(t) . \quad (53)$$

Let us now compute $S_2 x(t, \lambda_1, \lambda_2) = \|x \star \psi_{\lambda_1} \star \psi_{\lambda_2} \star \phi(t)\|$. If $T^{-1} \ll \lambda_1 Q_1^{-1}$ then (53) together with (48) shows that

$$\frac{S_2 x(t, \lambda_1, \lambda_2)}{S_1 x(t, \lambda_1)} = \frac{|a \star \psi_{\lambda_2} \star \phi(t)|}{a \star \phi(t)} + \tilde{\epsilon}(t) , \quad (54)$$

where

$$0 \leq \tilde{\epsilon}(t) \leq \frac{2|(a\epsilon) \star \psi_{\lambda_2} \star \phi(t)|}{\pi^{1/2} \|\psi_{\lambda_1}\| a \star \phi(t)} . \quad (55)$$

Observe that

$$E(|(a\epsilon) \star \psi_{\lambda_2} \star \phi(t)|) = E(|(a\epsilon) \star \psi_{\lambda_2}(t)|) \leq E(|(a\epsilon) \star \psi_{\lambda_2}(t)|^2)^{1/2} .$$

Lemma 1 applied to $z = \epsilon$ and $h = \psi_{\lambda_2}$ gives the following upper bound:

$$\mathbb{E}(|(a\epsilon) \star \psi_{\lambda_2}(t)|^2) \leq (1 - \pi/4) |a|^2 \star |\psi_{\lambda_2}|^2(t) . \quad (56)$$

One can write $|\psi_{\lambda_2}(t)| = \lambda_2 Q_2^{-1} \theta(\lambda_2 Q_2^{-1} t)$ where $\theta(t)$ satisfies $\int \theta(t) dt \sim 1$. Similarly to (52), we thus verify that if the square of the average of $a(t)$ is of the order of the average of $a^2(t)$ then

$$\frac{|a|^2 \star |\psi_{\lambda_2}|^2(t)}{|a \star \phi(t)|^2} \sim \frac{\lambda_2}{Q_2} . \quad (57)$$

Since $\|\psi_{\lambda_1}\|^2 \sim \lambda_1 Q_1^{-1}$, it results from (55,56,57) that $0 \leq \mathbb{E}(\tilde{\epsilon}(t)) \leq C (4/\pi - 1)^{1/2} (\lambda_2 Q_1)^{1/2} (\lambda_1 Q_2)^{-1/2}$ with $C \sim 1$.

References

- [1] J. Andén and S. Mallat, “Multiscale scattering for audio classification,” in *Proc. ISMIR*, Miami, Florida, Unites States, Oct. 24-28 2011, pp. 657–662.
- [2] L. Atlas and S. Shamma, “Joint acoustic and modulation frequency,” *EURASIP J. Appl. Signal Process.*, vol. 7, pp. 668–675, 2003.
- [3] E. Battenberg and D. Wessel, “Analyzing drum patterns using conditional deep belief networks,” in *Proc. ISMIR*, 2012.
- [4] C. Baugé, M. Lagrange, J. Andén, and S. Mallat, “Representing environmental sounds using the separable scattering transform,” in *Proc. IEEE ICASSP*, 2013.

- [5] J. Bruna and S. Mallat, “Invariant scattering convolution network,” *IEEE Trans. Pattern Anal. Mach. Intell.*, to appear, <http://arxiv.org/abs/1203.1513>.
- [6] J. Bruna, “Scattering representations for recognition,” Ph.D. dissertation, Ecole Polytechnique, 2013.
- [7] E. J. Candès, Y. Eldar, T. Strohmer, and V. Voroninski, “Phase retrieval via matrix completion,” *arXiv preprint arXiv:1109.0573*, 2011.
- [8] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [9] H.-A. Chang and J. R. Glass, “Hierarchical large-margin gaussian mixture models for phonetic classification,” in *Proc. IEEE ASRU*. IEEE, 2007, pp. 272–277.
- [10] X. Chen and P. J. Ramadge, “Music genre classification using multiscale scattering and sparse representations,” in *Proc. CISS*, 2013.
- [11] T. Chi, P. Ru, and S. Shamma, “Multiresolution spectrotemporal analysis of complex sounds,” *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.
- [12] V. Chudáček, J. Andén, S. Mallat, P. Abry, and M. Doret, “Scattering transform for intrapartum fetal heart rate characterization and acidosis detection,” in *Proc. IEEE EMBC*, 2013.
- [13] P. Clarkson and P. J. Moreno, “On the use of support vector machines for phonetic classification,” in *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 2. IEEE, 1999, pp. 585–588.
- [14] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.
- [15] T. Dau, B. Kollmeier, and A. Kohlrausch, “Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers,” *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2892–2905, 1997.
- [16] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.
- [17] D. Ellis, X. Zeng, and J. McDermott, “Classifying soundtracks with audio texture features,” in *Proc. IEEE ICASSP*, Prague, Czech Republic, May. 22-27 2001, pp. 5880–5883.

- [18] W. Fisher, G. Doddington, and K. Goudie-Marshall, “The darpa speech recognition research database: specifications and status,” in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.
- [19] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.
- [20] A. K. Halberstadt, “Heterogeneous acoustic measurements and multiple classifiers for speech recognition,” Ph.D. dissertation, Massachusetts Institute of Technology, 1998.
- [21] P. Hamel and D. Eck, “Learning features from music audio with deep belief networks,” in *Proc. ISMIR*, 2010.
- [22] M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. LeCun, “Unsupervised learning of sparse features for scalable audio classification,” in *Proc. ISMIR*, 2011.
- [23] H. Hermansky, “The modulation spectrum in the automatic recognition of speech,” in *Proc. IEEE ASRU*, 1997, pp. 140–147.
- [24] E. J. Humphrey, T. Cho, and J. P. Bello, “Learning a robust tonnetz-space transform for automatic chord recognition,” in *Proc. IEEE ICASSP*, 2012, pp. 453–456.
- [25] Y. LeCun, K. Kavukcuoglu, and C. Farabet, “Convolutional networks and applications in vision,” in *Proc. IEEE ISCAS*, 2010.
- [26] C. Lee, J. Shih, K. Yu, and H. Lin, “Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features,” *IEEE Transactions on Multimedia*, vol. 11, no. 4, pp. 670–682, 2009.
- [27] H. Lee, P. Pham, Y. Largman, , and A. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Proc. NIPS*, 2009.
- [28] K.-F. Lee and H.-W. Hon, “Speaker-independent phone recognition using hidden markov models,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 11, pp. 1641–1648, 1989.
- [29] L. Lucy, “An iterative technique for the rectification of observed distributions,” *Astron. J.*, vol. 79, p. 745, 1974.
- [30] S. Mallat, *A wavelet tour of signal processing*. Academic Press, 1999.
- [31] —, “Group invariant scattering,” *Commun. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.

- [32] J. McDermott and E. Simoncelli, “Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis,” *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.
- [33] N. Mesgarani, M. Slaney, and S. A. Shamma, “Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations,” *IEEE Audio, Speech, Language Process.*, vol. 14, no. 3, pp. 920–930, 2006.
- [34] A. Mohamed, G. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, 2012.
- [35] J. Nam, J. Herrera, M. Slaney, and J. Smith, “Learning sparse feature representations for music annotation and retrieval,” in *Proc. ISMIR*, 2012.
- [36] R. D. Patterson, “Auditory images: How complex sounds are represented in the auditory system,” *Journal of the Acoustical Society of Japan (E)*, vol. 21, no. 4, pp. 183–190, 2000.
- [37] M. Ramona and G. Peeters, “Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection,” in *Proc. IEEE ICASSP*, 2011, pp. 477–480.
- [38] T. N. Sainath, D. Nahamoo, D. Kanevsky, B. Ramabhadran, and P. Shah, “A convex hull approach to sparse representations for exemplar-based speech recognition,” in *Proc. IEEE ASRU*. IEEE, 2011, pp. 59–64.
- [39] L. Sifre and S. Mallat, “Rotation, scaling and deformation invariant scattering for texture discrimination,” in *Proc. CVPR*, 2013.
- [40] M. Slaney and R. Lyon, *Visual representations of speech signals*. M. Cooke, S. Beet and M. Crawford (Eds.) John Wiley and Sons, 1993, ch. On the importance of time—a temporal representation of sound, pp. 95–116.
- [41] E. C. Smith and M. S. Lewicki, “Efficient auditory coding,” *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.
- [42] S. Sukittanon, L. Atlas, and J. Pitton, “Modulation scale analysis for content identification,” *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 3023–3035, 2004.
- [43] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [44] I. Waldspurger, A. d’Aspremont, and S. Mallat, “Phase recovery, max-cut and complex semidefinite programming,” CMAP, Ecole Polytechnique, Tech. Rep., 2012.
- [45] I. Waldspurger and S. Mallat, “Recovering the phase of a complex wavelet transform,” CMAP, Ecole Polytechnique, Tech. Rep., 2012.