Deep Scattering Spectrum

Joakim Andén, *Student Member, IEEE,* Stéphane Mallat, *Fellow, IEEE*

***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.**

***Index Terms*—Audio classification, deep neural networks, MFCC, modulation spectrum, wavelets.**

I. 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 [1].

Spectrograms compute locally time-shift invariant descriptors over durations limited by a window. However, Section II shows that high-frequency spectrogram coefficients are not stable to variability due to time-warping deformations, which occur in most signals, particularly in audio. Stability means that small signal deformations produce small modifications of the representation, measured with a Euclidean norm. This is particularly important for classification. Mel-frequency spectograms are obtained by averaging spectrogram values over mel-frequency bands. It improves stability to time warping, but it also removes information. Over time intervals larger than 25ms, the information loss becomes too important, which is why mel-frequency spectrograms and MFCCs, are limited to such short time intervals. Modulation spectrum decompositions [2]–[10] 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 III shows that the information lost by mel-frequency spectrograms can be recovered with multiple layers of wavelet coefficients. In addition to being locally invariant to time-shifts, this representation is also stable to time-warping deformation. Known as a scattering transform [11], it is computed through a cascade of wavelet transforms and modulus nonlinearities. The computational structure is similar to a convolutional deep neural network [12]–[18], 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 [19], [20], also used for audio processing [21]. Its energy conservation and other mathematical properties are reviewed in Section IV. An approximate inverse scattering transform is introduced in Section V, with numerical examples. Section VI 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 VII 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 classification [22], [23], where spatial translation invariance is crucial. In audio, the analogous time-shift invariance is also important, but scattering transforms are computed with very different wavelets. They have a better frequency resolution, which is adapted to audio frequency structures. Section VIII explains how to adapt and optimize the 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 segment 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/data/scattering/*.

II. Mel-frequency Spectrum

Section II-A 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 II-B relates the mel-

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

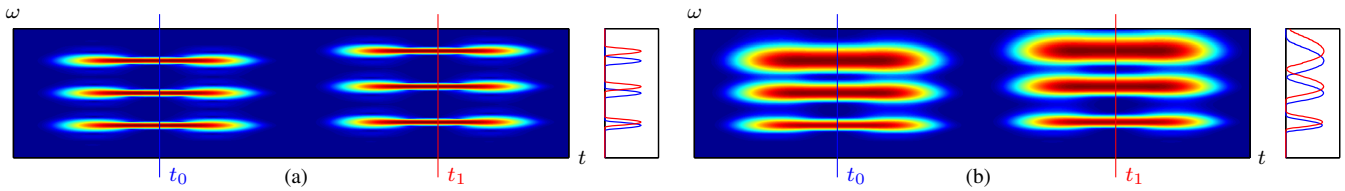

Fig. 1. (a) Spectrogram $\log|\widehat{x}(t,\omega)|$ for a harmonic signal $x(t)$ (centered in t_0) followed by $\log|\widehat{x}_\tau(t,\omega)|$ for $x_\tau(t)=x((1-\epsilon)t)$ (centered in t_1), as a function of t and ω. The right graph plots $\log|\widehat{x}(t_0,\omega)|$ (blue) and $\log|\widehat{x}_\tau(t_1,\omega)|$ (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.

frequency spectrogram to the amplitude output of a filter bank which computes a wavelet transform.

A. Fourier Invariance and Deformation Instability

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

$$|\widehat{x}_c(\omega)|=|\widehat{x}(\omega)|\ . \tag{1}$$

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

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

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

However, invariance to time-shifts is often not enough. 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 $x(t)$ and all τ with $\sup_t|\tau'(t)|<1$

$$\|\Phi(x)-\Phi(x_\tau)\|\le C\sup_t|\tau'(t)|\,\|x\|\ . \tag{3}$$

The constant C is a measure of stability.

This Lipschitz continuity property implies that time-warping deformations are locally linearized by $\Phi(x)$. Indeed, Lipschitz continuous operators are almost everywhere differentiable. It results that $\Phi(x)-\Phi(x_\tau)$ can be approximated by a linear operator if $\sup_t|\tau'(t)|$ is small. A family of small deformations thus generate a linear space. In the transformed space, an invariant to these deformations can then be computed with a linear projector on the orthogonal complement to this linear space. In Section VIII we use linear discriminant classifiers to become selectively invariant to small time-warping deformations.

A Fourier modulus representation $\Phi(x)=|\widehat{x}|$ 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

$$\||\widehat{x}|-|\widehat{x_\tau}|\|\le C\,\epsilon\,\|x\|\ . \tag{4}$$

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 component at 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

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

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 $n\epsilon\xi$ becomes larger than the bandwidth of $\widehat{g}$. Consequently, the harmonics $\widehat{g}(\omega(1-\epsilon)^{-1}-n\xi)$ of $\widehat{x}_\tau$ do not overlap with the harmonics $\widehat{g}(\omega-n\xi)$ of $\widehat{x}$. 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 deformation stability condition (4) is not satisfied for any $C>0$.

The autocorrelation $Rx(u)=\int x(t)\,x^\star(t-u)\,dt$ 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 $\|Rx-Rx_\tau\|=(2\pi)^{-1}\||\widehat{x}|^2-|\widehat{x}_\tau|^2\|$.

B. 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 $\widehat{\psi}_\lambda(\omega)$:

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

The band-pass filters $\widehat{\psi}_\lambda$ have a constant-Q frequency bandwidth at high frequencies. Their frequency support is centered at λ with a bandwidth of the order of λ/Q. At lower frequencies, instead of being constant-Q, the bandwidth of $\widehat{\psi}_\lambda$ remains equal to $2\pi/T$.

The mel-frequency averaging removes deformation instability created by large displacements of high frequencies under dilations. If $x_\tau(t)=x((1-\epsilon)t)$ then we saw that each frequency component at ω_0 is moved by $\epsilon|\omega_0|$, which may be large if $|\omega_0|$ is large. However, the mel-scale filter $\widehat{\psi}_\lambda(\omega)$ covering the frequency ω_0 has a frequency bandwidth of the order of $\lambda/Q\sim|\omega_0|/Q$. As a result, the relative error after averaging by $|\widehat{\psi}|^2$ is of the order of ϵQ. This is illustrated by Figure 1(b) on a harmonic signal x. After mel-frequency averaging, the frequency partials of x and x_τ overlap at all frequencies. One can verify that $\|Mx(t,\lambda)-M_\tau x(t,\lambda)\|\le C\,\epsilon\|x\|$, where C is proportional to Q, and does not depend upon ϵ and x.

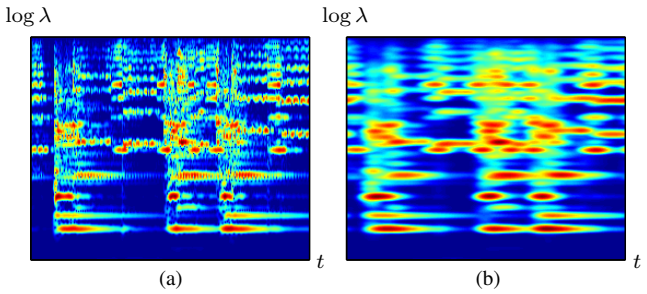

Fig. 2. (a): Scalogram $\log|x\star\psi_\lambda(t)|^2$ for a musical signal, as a function of t and λ. (b): Averaged scalogram $\log|x\star\psi_\lambda|^2\star\phi^2(t)$ with a lowpass filter ϕ of duration $T=190$ms.

Unlike the spectrogram (2), the mel-frequency spectrogram (6) satisfies the Lipschitz deformation stability condition (3).

Mel-scale averaging provides time-warping stability but loses information. We show that this frequency averaging is equivalent to a time averaging of a filter bank output, which will provide a strategy to recover the lost information. 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

$$\mathrm{M}x(t,\lambda)=\frac{1}{2\pi}\int|\widehat{x_t}(\omega)|^2\,|\widehat{\psi_\lambda}(\omega)|^2\,d\omega \tag{7}$$

$$=\int|x_t\star\psi_\lambda(v)|^2\,dv \tag{8}$$

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

If $\lambda\gg Q/T$ 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

$$\mathrm{M}x(t,\lambda)\approx\int\left|\int x(u)\psi_\lambda(v-u)du\right|^2|\phi(v-t)|^2dv \tag{10}$$

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

The frequency averaging of the spectrogram is thus nearly equal to the time averaging of $|x\star\psi_\lambda|^2$. In this formulation, the window ϕ acts as a lowpass filter, ensuring that the representation is locally invariant to time-shifts smaller than T. Section III-A 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 $T=190$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 amplitude modulations of $|x\star\psi_\lambda(t)|$ at scales smaller than T, which are important in audio perception. The spectrum of these modulation envelopes can be computed from the spectrogram [2]–[5] of $|x\star\psi_\lambda|$, or represented with a short-time autocorrelation [6], [7]. 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 II-A showed that spectrograms and autocorrelations have deformation instabilities. Constant-Q averaged modulation spectra [9], [10] 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.

III. Wavelet Scattering Transform

A scattering transform recovers the information lost by a mel-frequency averaging with a cascade of wavelet decompositions and modulus operators [11]. 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 III-B.

A. 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 $\widehat{\psi}(0)=0$. We consider complex wavelets with quadrature phase such that $\widehat{\psi}(\omega)\approx 0$ for $\omega<0$. For any $\lambda>0$, a dilated wavelet of center frequency λ is written

$$\psi_\lambda(t)=\lambda\,\psi(\lambda\,t)\ \text{ and hence }\ \widehat{\psi}_\lambda(\omega)=\widehat{\psi}\Big(\frac{\omega}{\lambda}\Big)\ . \tag{12}$$

The center frequency of $\widehat{\psi}$ 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 $\widehat{\psi}$ is of the order of Q^{-1}, to cover the whole frequency axis with these band-pass wavelet filters. The support of $\widehat{\psi}_\lambda(\omega)$ 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 $\widehat{\psi}_\lambda$ 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 $\lambda\in\Lambda$:

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

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

$$A(\omega)=|\widehat{\phi}(\omega)|^2+\frac{1}{2}\sum_{\lambda\in\Lambda}\Big(|\widehat{\psi}_\lambda(\omega)|^2+|\widehat{\psi}_\lambda(-\omega)|^2\Big) \tag{14}$$

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

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

This condition implies that the wavelet transform W is a stable and invertible operator. Multiplying (15) by $|\widehat{x}(\omega)|^2$ and applying the Plancherel formula [24] gives

$$(1-\alpha)\|x\|^2 \leq \|Wx\|^2 \leq \|x\|^2 \ , \tag{16}$$

where $\|x\|^2 = \int |x(t)|^2 dt$ and where 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

$$x(t) = (x \star \phi) \star \overline{\phi}(t) + \sum_{\lambda \in \Lambda} \text{Real}\Big((x \star \psi_\lambda) \star \overline{\psi}_\lambda(t) \Big) \ , \tag{17}$$

with reconstruction filters defined by

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

where z^* is the complex conjugate of $z \in \mathbb{C}$. If $\alpha = 0$ in (15) then W is said to be a tight frame operator, in which case $\overline{\phi}(t) = \phi(-t)$ and $\overline{\psi}_\lambda(t) = \psi^*_\lambda(-t)$.

One may define an analytic wavelet with an octave resolution Q as $\psi(t) = e^{it}\,\theta(t)$ and hence $\widehat{\psi}(\omega) = \widehat{\theta}(\omega - 1)$ where $\widehat{\theta}$ is the transfer function of a low-pass filter whose bandwidth is of the order of Q^{-1}. If $\widehat{\theta}(-1) \neq 0$ then we define $\widehat{\psi}(\omega) = \widehat{\theta}(\omega-1) - \widehat{\theta}(\omega)\widehat{\theta}(-1)/\widehat{\theta}(0)$, which guarantees that $\widehat{\psi}(0) = 0$. If θ is a Gaussian then ψ is called a Morlet wavelet, which is almost analytic because $|\widehat{\psi}(\omega)|$ is small but not strictly zero for $\omega < 0$. Figure 3 shows Morlet wavelets $\widehat{\psi}_\lambda$ with $Q = 8$. In this case ϕ is also chosen to be a Gaussian. For $Q = 1$, tight frame 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 [11]. Unless indicated otherwise, wavelets used in this paper are Morlet wavelets.

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

$$|W|x = \Big(x \star \phi(t) \, , \, |x \star \psi_\lambda(t)| \Big)_{t \in \mathbb{R}, \lambda \in \Lambda} \ . \tag{19}$$

Taking the modulus of analytic wavelet coefficient can be interpreted as a sub-band Hilbert envelope demodulation. Demodulation is used to separate carriers and modulation envelopes. When a carrier or pitch frequency can be detected, then a linear coherent demodulation is efficiently implemented by multiplying the analytic signal with the conjugate of the detected carrier [25]–[27]. However, many signals such as unvoiced speech are not modulated by isolated carrier frequency, in which case coherent demodulation is not well defined. Nonlinear Hilbert envelope demodulations apply to any band-pass analytic signals, but if a carrier is present then the Hilbert envelope depends both on the carrier and on the amplitude modulation. Section VI-C explains how to isolate amplitude modulation coefficients from Hilbert envelope measurements, whether a carrier is present or not.

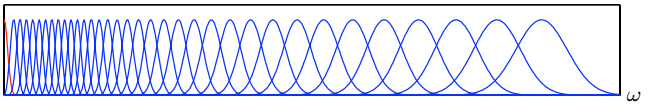

Fig. 3. Morlet wavelets $\widehat{\psi}_\lambda(\omega)$ with $Q = 8$ wavelets per octave, for different λ. The low frequency filter $\widehat{\phi}(\omega)$ (in red) is a Gaussian.

Although a wavelet modulus operator removes the complex phase, it does not lose information because the temporal variation of the multiscale envelopes is kept. A signal cannot be reconstructed from the modulus of its Fourier transform, but it can be recovered from the modulus of its wavelet transform. Since the time variable t is not subsampled, a wavelet transform has more coefficients than the original signal. These coefficients are highly redundant when filters have a significant frequency overlap. For particular families of analytic wavelets, one can prove that $|W|$ is an invertible operator with a continuous inverse [28]. This is further studied in Section V.

The operator $|W|$ is 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

$$\| \, |W|x - |W|x'\|^2 \leq \|Wx - Wx'\|^2 \leq \|x - x'\|^2 \ .$$

If W is a tight frame operator then $\| \, |W|x\| = \|Wx\| = \|x\|$ so $|W|$ preserves the signal norm.

B. Deep Scattering Network

We showed in (11) that mel-frequency spectral coefficients $Mx(t, \lambda)$ are approximately equal to averaged squared wavelet coefficients $|x \star \psi_\lambda|^2 \star |\phi|^2(t)$. Large wavelet coefficients are considerably amplified by the square operator. To avoid amplifying outliers, we remove the square and calculate $|x \star \psi_\lambda| \star \phi(t)$ instead. High frequencies removed by the low-pass filter ϕ are recovered by a new set of wavelet modulus coefficients. Cascading this procedure defines a scattering transform.

A locally translation invariant descriptors of x is obtained with a time-average $S_0 x(t) = x \star \phi(t)$, which removes all high frequencies. These high-frequencies are recovered by a wavelet modulus transform

$$|W_1|x = \Big(x \star \phi(t) \, , \, |x \star \psi_{\lambda_1}(t)| \Big)_{t \in \mathbb{R}, \lambda_1 \in \Lambda_1} \ .$$

It is 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 $S_0 x(t) \approx 0$. Approximate mel-frequency spectral coefficients are obtained by averaging the wavelet modulus coefficients with ϕ:

$$S_1 x(t, \lambda_1) = |x \star \psi_{\lambda_1}| \star \phi(t) \ . \tag{20}$$

These are called first-order scattering coefficients. They are computed with a second wavelet modulus transform $|W_2|$

applied to each $|x \star \psi_{\lambda_1}|$, which also provides complementary high frequency wavelet coefficients:

$$|W_2|\,|x \star \psi_{\lambda_1}| = \Big(|x \star \psi_{\lambda_1}| \star \phi\,,\, ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|\Big)_{\lambda_2 \in \Lambda_2}.$$

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 information over as few wavelet coefficients as possible. These coefficients are averaged by the lowpass filter ϕ of size T, which ensures local invariance to time-shifts, as with the first-order coefficients. It defines second-order scattering coefficients:

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

These averages are computed by applying a third wavelet modulus transform $|W_3|$ to each $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$. It computes their wavelet modulus 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:

$$U_m x(t, \lambda_1, ..., \lambda_m) = |\,||x \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}(t)|\ , \qquad (21)$$

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

$$\begin{aligned} S_m x(t, \lambda_1, ..., \lambda_m) &= |\,||x \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}| \star \phi(t) \\ &= U_m x(., \lambda_1, ..., \lambda_m) \star \phi(t)\ . \end{aligned}$$

Applying $|W_{m+1}|$ on $U_m x$ computes both $S_m x$ and $U_{m+1} x$:

$$|W_{m+1}|\,U_m x = (S_m x\,,\, U_{m+1} x)\ . \qquad (22)$$

A scattering decomposition of maximal order l is thus defined by initializing $U_0 x = x$, and recursively computing (22) for $0 \leq m \leq l$. This scattering transform is illustrated in Figure 4. The final scattering vector aggregates all scattering coefficients for $0 \leq m \leq l$:

$$Sx = (S_m x)_{0 \leq m \leq l}. \qquad (23)$$

The scattering cascade of convolutions and non-linearities can also be interpreted as a convolutional network [12], where $U_m x$ is the set of coefficients of the mth internal network layer. These networks have been shown to be highly effective for audio classification [13]–[18]. However, unlike standard convolutional networks, each such layer has an output $S_m x = U_m x \star \phi$, not just the last layer. In addition, all filters are predefined wavelets and are not learned from training data. A scattering transform, like MFCCs, provide a low-level invariant representation of the signal, without learning. It relies on prior information concerning the type of invariants that need to be computed, in this case relatively to time-shifts and time-warping deformations, or in Section VII relatively to frequency transpositions. When no such information is available, or if the sources of variabilities are much more complex, then it is necessary to learn them from examples, which can be done with deep neural networks [13]–[18]. In that sense both approaches are complementary.

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 V. Sparsity seems also to play an important role for classification [29], [30]. For audio signals x, choosing $Q_1 = 8$ wavelets per octave has been shown to provide sparse representations of a mix of speech, music and environmental signals [31]. It nearly corresponds to a mel-scale frequency subdivision.

At the second order, choosing $Q_2 = 1$ defines wavelets with more narrow time support, which are better adapted to characterize transients and attacks. Section VI 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 always set $Q_m = 1$, but we shall 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 [19], [20]. The first filter bank with $Q_1 = 8$ models the cochlear filtering, whereas the second filter bank corresponds to later processing in the models with filters that have $Q_2 = 1$ [19], [20].

IV. Scattering Properties

We briefly review important properties of scattering transforms, including stability to time-warping deformation, energy conservation, and describe a fast computational algorithm.

A. Time-Warping Stability

Stability to time-warping allows one to use linear operators for calculating invariant descriptors to small time-warping deformations. 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 II explains that mel-frequency spectrograms become stable to time-warping deformation with a frequency averaging. One can prove that a scattering representation $\Phi(x) = Sx$ satisfies the Lipschitz continuity condition (3) because wavelets are stable to time-warping [11]. Let us write $\psi_{\lambda,\tau}(t) = \psi_\lambda(t - \tau(t))$. One can verify that there exists $C > 0$ such that $\|\psi_\lambda - \psi_{\lambda,\tau}\| \leq C \|\psi_\lambda\| \sup_t |\tau'(t)|$, for all λ and all $\tau(t)$. This property is at the core of the scattering stability to time-warping deformations.

The squared Euclidean norm of a scattering vector Sx is the sum of its coefficients squared at all orders:

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

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 support of ϕ. One can

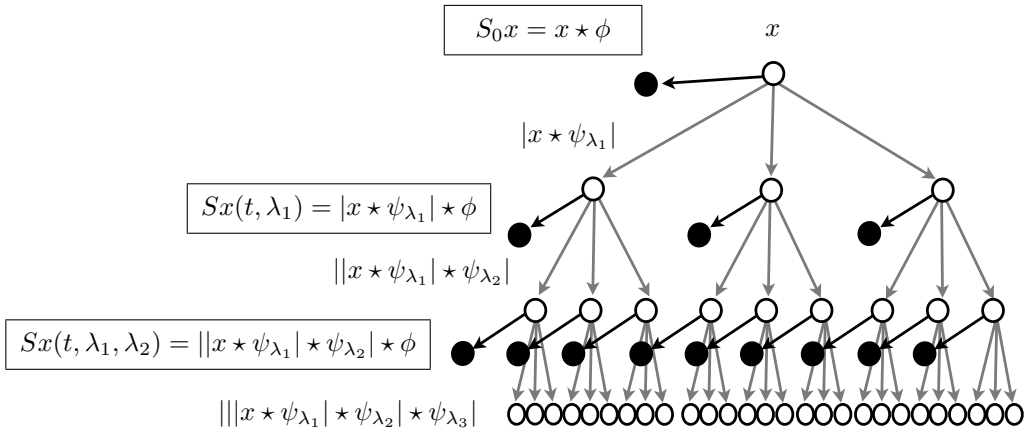

Fig. 4. A scattering transform iterates on wavelet modulus operators $|W_m|$ to compute cascades of m wavelet convolutions and moduli stored in $U_m x$, and to output averaged scattering coefficients $S_m x$.

prove that there exists a constant C such that for all x and any such τ [11]:

$$\|Sx_\tau - Sx\| \leq C \sup_t |\tau'(t)| \, \|x\| \ , \tag{24}$$

up to second-order terms. As explained for mel-spectral decompositions, the constant C is inversely proportional to the octave bandwidth of wavelet filters. Over multiple scattering layers, we get $C = C_0(\max_m Q_m)^{-1}$. For Morlet wavelets numerical experiments on many examples give $C_0 \approx 2$.

T	$m=0$	$m=1$	$m=2$	$m=3$
23ms	0.0%	94.5%	4.8%	0.2%
93ms	0.0%	68.0%	29.0%	1.9%
370ms	0.0%	34.9%	53.3%	11.6%
1.5 s	0.0%	27.7%	56.1%	24.7%

TABLE I
AVERAGED VALUES $\|S_m x\|^2/\|x\|^2$ COMPUTED FOR SIGNALS x IN THE TIMIT SPEECH DATASET [32], AS A FUNCTION OF ORDER m AND AVERAGING SCALE T. FOR $m=1$, $S_m x$ IS CALCULATED BY MORLET WAVELETS WITH $Q_1 = 8$, AND FOR $m = 2, 3$ BY CUBIC SPLINE WAVELETS WITH $Q_2 = Q_3 = 1$.

B. Contraction and Energy Conservation

We show that a scattering transform is contractive and can preserve energy. We denote $\|Ax\|^2$ the squared Euclidean norm of a vector of coefficients Ax, such as $W_m x$, $S_m x$, $U_m x$ or Sx. Since Sx is computed by cascading wavelet modulus operators $|W_m|$, which are all contractive, it results that S is also contractive:

$$\|Sx - Sx'\| \leq \|x - x'\| \ . \tag{25}$$

A scattering transform is therefore stable to additive noise.

If each wavelet transform is a tight frame, that is $\alpha = 0$ in (15), each $|W_m|$ preserves the signal norm. Applying this property to $|W_{m+1}|U_m x = (S_m x \, , \, U_{m+1} x)$ yields

$$\|U_m x\|^2 = \|S_m x\|^2 + \|U_{m+1} x\|^2 \ . \tag{26}$$

Summing these equations $0 \leq m \leq l$ proves that

$$\|x\|^2 = \|Sx\|^2 + \|U_{l+1} x\|^2 \ . \tag{27}$$

Under appropriate assumptions on the mother wavelet ψ, one can prove that $\|U_{l+1} x\|$ goes to zero as l increases, which implies that $\|Sx\| = \|x\|$ for $l = \infty$ [11]. 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 $S_m x = U_m x \star \phi$.

One can verify numerically that $\|U_{l+1} x\|$ converges to zero exponentially when l goes to infinity and hence that $\|Sx\|$ converges exponentially to $\|x\|$. Table I 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 53% of the signal energy is captured in $S_2 x$. Section VI 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 12% 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 are less important than first- and second-order coefficients. We therefore concentrate on second-order scattering representations:

$$Sx = \Big(S_0 x(t) \, , \, S_1 x(t, \lambda_1) \, , \, S_2 x(t, \lambda_1, \lambda_2) \Big)_{t, \lambda_1, \lambda_2} \ . \tag{28}$$

C. Fast Scattering Computation

Subsampling scattering vectors provide a reduced representation, which leads to a faster implementation. Since the averaging window ϕ has a duration of the order of T, we compute scattering vectors with half-overlapping windows at $t = kT/2$ with $k \in \mathbb{Z}$.

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 wavelets ψ_{λ_1} 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 $S_2 x(t, \lambda_1, \lambda_2)$ which needs to be computed 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 approximately the same frequency bandwidth as $\widehat{\psi}_{\lambda_1}$. 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 $\widehat{\psi}_{\lambda_2}$ 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

$$\lambda_2 \leq \max(\lambda_1 Q_1^{-1}, 2\pi T^{-1}) \ . \tag{29}$$

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 mth-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, with the above subsampling we compute approximately $2N$ first-order wavelet coefficients $\{|x \star \psi_{\lambda_1}(t)|\}_{t, \lambda_1 \in \Lambda_1}$. Similarly, $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t)|$ is subsampled in time at a rate twice its bandwidth. Over the same time frame, the total number of second-order wavelet coefficients for all t, λ_1 and λ_2 stays below $2N$. With a fast Fourier transform (FFT), these first- and second-order wavelet modulus coefficients are computed using $O(N \log N)$ operations. The resulting scattering coefficients $S_1 x(t, \lambda_1)$ and $S_2 x(t, \lambda_1, \lambda_2)$ are also calculated with $O(N \log N)$ operations, with FFT convolutions with ϕ.

V. 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 $|W_m|$, the inversion approximately inverts each $|W_m|$ for $m < l$. At the maximum depth $m = l$, the algorithm begins with a deconvolution, estimating $U_l x(t)$ at all t on the sampling grid of $x(t)$, from $S_l x(kT/2) = U_l x \star \phi(kT/2)$.

Because of the subsampling, one cannot compute $U_l x$ from $S_l x$ exactly. This deconvolution is thus the main source of error. To take advantage of the fact that $U_l x \geq 0$, the deconvolution is computed with the Richardson-Lucy algorithm [33], which preserves positivity if $\phi \geq 0$. We initialize $y_0(t)$ by interpolating $S_l x(kT/2)$ linearly on the sampling grid of x, which introduces error because of aliasing. The Richardson-Lucy deconvolution iteratively computes

$$y_{n+1}(t) = y_n(t) \cdot \left[\left(\frac{y_0}{y_n \star \phi} \right) \star \widetilde{\phi}(t) \right] \ , \tag{30}$$

with $\widetilde{\phi}(t) = \phi(-t)$. Since it converges to the pseudo-inverse of the convolution operator applied to y_0, it 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. The result is then our estimate of $U_l x$.

Once an estimation of $U_l x$ is calculated by deconvolution, we compute an estimate $\widetilde{x}$ of x by inverting each $|W_m|$ for $l \geq m > 0$. The wavelet transform of a signal x of size N is a vector $Wx = (x \star \phi, x \star \psi_\lambda)_{\lambda \in \Lambda}$ of about $QN \log_2 N$ coefficients, where Q is the number of wavelets ψ_λ per octave. These coefficients live in a subspace $\mathbf{V}$ of dimension N. To recover Wx from $|W|x = (x \star \phi, |x \star \psi_\lambda|)_{\lambda \in \Lambda}$, we search for a vector in $\mathbf{V}$ whose modulus values are specified by $|W|x$. This a non-convex optimization problem. Recent convex relaxation approaches [34], [35] 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. The inversion of $|W|$ is typically more stable when $|W|x$ is sparse because there is no phase to recover if $|x \star \psi_\lambda| = 0$. This motivates using wavelets ψ_{λ_m} which provide sparse representations at each order m.

Griffin & Lim [36] showed that alternating projections recovers good quality audio signals from spectrogram values, but with large mean-square errors because the algorithm is trapped in local minima. The same algorithm inverts $|W|$ by alternating projections on the wavelet transform space $\mathbf{V}$ and on the modulus constraints. An estimation $\widetilde{x}$ of x is calculated from $|W|x$, by initializing $\widetilde{x}_0$ to be a Gaussian white noise. For any $n \geq 0$, $\widetilde{x}_{n+1}$ is computed from $\widetilde{x}_n$ by first adjusting the modulus of its wavelet coefficients, with a non-linear projector

$$z_\lambda(t) = |x \star \psi_\lambda(t)| \, \frac{\widetilde{x}_n \star \psi_\lambda(t)}{|\widetilde{x}_n \star \psi_\lambda(t)|} \ . \tag{31}$$

Applying the wavelet transform pseudo-inverse (17) yields

$$\widetilde{x}_{n+1} = x \star \phi \star \overline{\phi}(t) + \sum_{\lambda \in \Lambda} \mathrm{Real}\Big(z_\lambda \star \overline{\psi}_\lambda(t) \Big) \ . \tag{32}$$

The dual filters are defined in (18). One can verify that $W\widetilde{x}_{n+1}$ is the orthogonal projection of $\{x \star \phi, z_\lambda\}_{\lambda \in \Lambda}$ in $\mathbf{V}$. Numerical experiments are performed with $n = 30$ iterations, and we set $\widetilde{x} = \widetilde{x}_n$.

When $l = 1$, an approximation $\widetilde{x}$ of x is computed from from $(S_0 x, S_1 x)$ by first estimating $U_1 x$ from $S_1 x = U_1 x \star \phi$ with the Richardson-Lucy deconvolution algorithm. We then compute $\widetilde{x}$ from $S_0 x$ and this estimation of $U_1 x$ by approximately inverting $|W_1|$ with the Griffin & Lim algorithm. When T is above $100\mathrm{ms}$, the deconvolution loses too much information, and audio reconstructions obtained from first-order coefficients are crude. Figure 5(a) shows the scalograms $\log |x \star \psi_{\lambda_1}(t)|$ of a speech and a music signal, and the

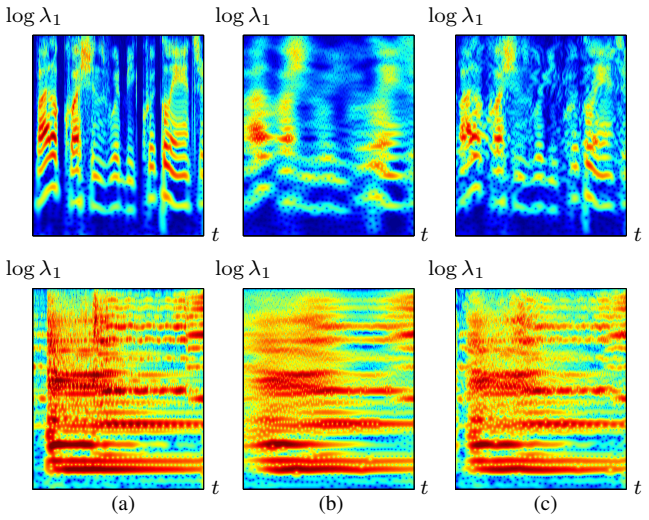

Fig. 5. *(a): Scalogram* $\log|x\star\psi_{\lambda_1}(t)|$ *for recordings of speech (top) and a cello (bottom). (b,c): Scalograms* $\log|\widetilde{x}\star\psi_{\lambda_1}(t)|$ *of reconstructions* $\widetilde{x}$ *from first-order scattering coefficients* ($l=1$) *in (b), and from first- and second-order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

scalograms $\log|\widetilde{x}\star\psi_{\lambda_1}(t)|$ of their approximations $\widetilde{x}$ from first-order scattering coefficients.

When $l=2$, the approximation $\widetilde{x}$ 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 $|W_2|$ and $|W_1|$ with the Griffin & Lim algorithm. Figure 5(c) shows $\log|\widetilde{x}\star\psi_{\lambda_1}(t)|$ for the same speech and music signals. Amplitude modulations, vibratos and attacks are restored with greater precision by incorporating second-order coefficients, yielding much better audio quality compared to first-order reconstructions. However, even with $l=2$, reconstructions become crude for $T\geq 500$ms. Indeed, the number of second-order scattering coefficients $Q_1Q_2\log_2^2 N/2$ 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/data/scattering/audio/*.

VI. Normalized Scattering Spectrum

To reduce redundancy and increase invariance, Section VI-A normalizes scattering coefficients. Section VI-B shows that normalized second order coefficients provide high-resolution spectral information through interferences. Section VI-C also proves that they characterize amplitude modulations of audio signals.

A. Normalized Scattering Transform

Scattering coefficients are renormalized to increase their invariance. It also decorrelates these coefficients at different orders. First order scattering coefficients are renormalized so that they become insensitive to multiplicative constants:

$$\widetilde{S}_1x(t,\lambda_1)=\frac{S_1x(t,\lambda_1)}{\sum_{\lambda\in\Lambda_1}S_1x(t,\lambda)+\epsilon}\ . \tag{33}$$

The constant ϵ is a silence detection threshold so that $\widetilde{S}_1x=0$ if $x=0$, which may be set to 0.

At any order $m\geq 2$, scattering coefficients are renormalized by previous order coefficients:

$$\widetilde{S}_mx(t,\lambda_1,...,\lambda_{m-1},\lambda_m)=\frac{S_mx(t,\lambda_1,...,\lambda_{m-1},\lambda_m)}{S_{m-1}x(t,\lambda_1,...,\lambda_{m-1})+\epsilon}\ . \tag{34}$$

A normalized scattering representation is defined by $\widetilde{S}x=(\widetilde{S}_mx)_{1\leq m\leq l}$. We shall mostly limit ourself to $l=2$.

For $m=2$,

$$\widetilde{S}_2x(t,\lambda_1,\lambda_2)=\frac{S_2x(t,\lambda_1,\lambda_2)}{S_1x(t,\lambda_1)+\epsilon}\ . \tag{35}$$

Let us show that these coefficients are nearly invariant to convolutions with filters h such that

$$\int|t|\,|h(t)|\,dt\ll\frac{\lambda_1}{Q_1}. \tag{36}$$

It results from this property that $\widehat{h}(\omega)$ is approximately constant on the support of $\widehat{\psi}_{\lambda_1}$, so $h\star\psi_{\lambda_1}(t)\approx\widehat{h}(\lambda_1)\psi_{\lambda_1}(t)$. Consequently, $|(x\star h)\star\psi_{\lambda_1}(t)|\approx|\widehat{h}(\lambda_1)|\,|x\star\psi_{\lambda_1}(t)|$, and

$$S_1(x\star h)(t,\lambda_1)\approx|\widehat{h}(\lambda_1)|S_1x(t,\lambda_1). \tag{37}$$

First-order scattering coefficients thus carry information about $\hat{h}$, but since

$$S_2(x\star h)(t,\lambda_1,\lambda_2)\approx|\widehat{h}(\lambda_1)|S_2x(t,\lambda_1,\lambda_2), \tag{38}$$

it results that $\widetilde{S}_2(x\star h)(t,\lambda_1,\lambda_2)\approx\widetilde{S}_2x(t,\lambda_1,\lambda_2)$. Normalized second-order coefficients are thus invariant to filtering by h. One can verify that this remains valid at any order $m\geq 2$.

B. 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.

Suppose x has two frequency components in the support of $\widehat{\psi}_{\lambda_1}$. We then have

$$x\star\psi_{\lambda_1}(t)=\alpha_1\,e^{i\xi_1t}+\alpha_2\,e^{i\xi_2t},$$

whose modulus squared equals

$$|x\star\psi_{\lambda_1}(t)|^2=|\alpha_1|^2+|\alpha_2|^2+2|\alpha_1\alpha_2|\cos(\xi_1-\xi_2)t. \tag{39}$$

We approximate $|x\star\psi_{\lambda_1}(t)|$ with a first-order expansion of the square root, which yields

$$|x\star\psi_{\lambda_1}(t)|\approx\sqrt{|\alpha_1|^2+|\alpha_2|^2}+\frac{|\alpha_1\alpha_2|}{\sqrt{|\alpha_1|^2+|\alpha_2|^2}}\cos(\xi_1-\xi_2)t. \tag{40}$$

If ϕ has a support of size $T\gg|\xi_1-\xi_2|$, then $S_1x(t,\lambda_1)\approx\sqrt{|\alpha_1|^2+|\alpha_2|^2}$, so

$$\widetilde{S}_2(t,\lambda_1,\lambda_2)=\frac{S_2x(t,\lambda_1,\lambda_2)}{S_1x(t,\lambda_1)}\approx|\widehat{\psi}_{\lambda_2}(\xi_2-\xi_1)|\,\frac{|\alpha_1\,\alpha_2|}{|\alpha_1|^2+|\alpha_2|^2}\ . \tag{41}$$

These normalized second-order coefficients are thus non-negligible when λ_2 is of the order of the frequency interval $|\xi_2-\xi_1|$. This shows that although the first wavelet $\widehat{\psi}_{\lambda_1}$ 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. The normalized second-order scattering coefficients (41) are large only if α_1 and α_2 have the same order of magnitude. This also conforms to auditory perception where a frequency interval is perceived only when the two frequency components have a comparable amplitude.

If $x \star \psi_{\lambda_1}(t) = \sum_n \alpha_n \, e^{i\xi_n t}$ has more frequency components, we verify similarly that $\widetilde{S}_2 x(t, \lambda_1, \lambda_2)$ 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 $\widehat{\psi}_{\lambda_1}$. 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 $\widetilde{S}_3 x(t, \lambda_1, \lambda_2, \lambda_3)$ 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 $\xi_1 = 600$Hz and $\xi_2 = 675$Hz, followed by an arpeggio of the same two notes. First-order coefficients $\log \widetilde{S}_1 x(t, \lambda_1)$ 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 $\log \widetilde{S}_2 x(t, \lambda_1, \lambda_2)$ for $\lambda_1 \approx \xi_1 = 600$Hz, as a function of λ_2. The chord creates large amplitude coefficients for $\lambda_2 = |\xi_2 - \xi_1| = 75$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.

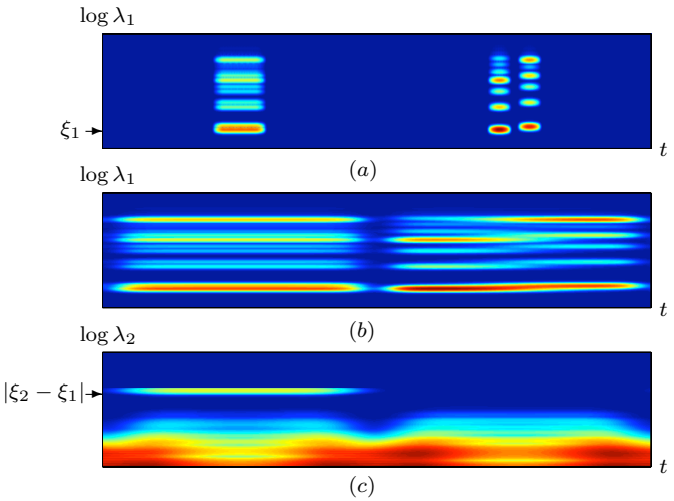

Fig. 6. *(a): Scalogram $\log |x \star \psi_{\lambda_1}(t)|$ for a signal with two notes, of fundamental frequencies $\xi_1 = 600$Hz and $\xi_2 = 675$Hz, first played as a chord and then as an arpeggio. (b): First-order normalized scattering coefficients $\log \widetilde{S}_1 x(t, \lambda_1)$ for $T = 512$ms. (c): Second-order normalized scattering coefficients $\log \widetilde{S}_2(t, \xi_1, \lambda_2)$ with $\lambda_1 = \xi_1$ as a function of t and λ_2. The chord interferences produce large coefficients for $\lambda_2 = |\xi_2 - \xi_1|$.*

C. 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, we show that amplitude modulations are characterized by normalized 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

$$x(t) = a(t)\,(e \star h)(t) \;. \tag{42}$$

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

$$\left(\int |t|\,|h(t)|\,dt \right)^{-1} \gg \frac{\lambda_1}{Q_1} \gg \sup_t |a'(t)| \;, \tag{43}$$

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

$$|x \star \psi_{\lambda_1}(t)| \approx |\widehat{h}(\lambda_1)|\,|e \star \psi_{\lambda_1}(t)|\,a(t) \;. \tag{44}$$

We 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:

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

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 (44) that

$$|x \star \psi_{\lambda_1}(t)| \approx |\widehat{h}(\lambda_1)|\,|\widehat{\psi}_{\lambda_1}(k\xi)|\,a(t) \;, \tag{45}$$

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

$$S_1 x(t, \lambda_1) \approx |\widehat{h}(\lambda_1)|\,|\widehat{\psi}_{\lambda_1}(k\xi)|\,a \star \phi(t) \;. \tag{46}$$

After the normalization (33), the coefficients $\widetilde{S}_1 x(t, \lambda_1)$ do not depend upon t and $a(t)$, as long as $a(t)$ is not negligible.

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 $\xi = 600$ Hz. Figure 7(b) shows that $\log \widetilde{S}_1 x(t, \lambda_1)$ has a harmonic structure, with an amplitude depending on $\log |\widehat{h}(\lambda_1)|$. The averaging by ϕ and the normalization remove the effect of the different modulation amplitudes $a(t)$ of these three voiced sounds.

Using (45), we compute second-order scattering coefficients

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

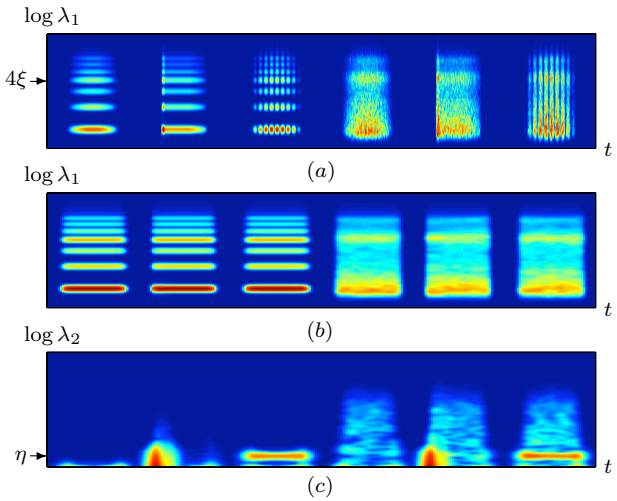

Fig. 7. *(a):* $\log|x \star \psi_{\lambda_1}(t)|$ *for a signal with three voiced sounds of same pitch* $\xi = 600$Hz *and same* $h(t)$ *but different amplitude modulations* $a(t)$*: first a smooth attack, then a sharp attack, then a tremolo of frequency* η*. It is followed by three unvoiced sounds created with the same* $h(t)$ *and same amplitude modulations* $a(t)$ *as the first three voiced sounds. (b): First-order scattering* $\log \widetilde{S}_1 x(t, \lambda_1)$ *with* $T = 128$ms*. (c): Second-order scattering* $\log \widetilde{S}_2 x(t, \lambda_1, \lambda_2)$ *displayed for* $\lambda_1 = 4\xi$*, as a function of* t *and* λ_2*.*

which implies that

$$\widetilde{S}_2 x(t, \lambda_1, \lambda_2) = \frac{S_2 x(t, \lambda_1, \lambda_2)}{S_1 x(t, \lambda_1)} \approx \frac{|a \star \psi_{\lambda_2}| \star \phi(t)}{a \star \phi(t)} \,. \quad (47)$$

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

Figure 7(c) displays $\log \widetilde{S}_2(t, \lambda_1, \lambda_2)$ 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 (47), 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 $\lambda_1 Q_1^{-1} \gg T^{-1}$, over frames where the signal energy is not negligible, Appendix A shows

$$\widetilde{S}_1 x(t, \lambda_1) \approx \frac{{\lambda_1}^{1/2} \, |\widehat{h}(\lambda_1)|}{\sum_{\lambda \in \Lambda_1} \lambda^{1/2} \, |\widehat{h}(\lambda)|} \quad (48)$$

First-order scattering coefficients are again proportional to $|\widehat{h}(\lambda_1)|$ 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 $e(t)$.

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

$$\widetilde{S}_2 x(t, \lambda_1, \lambda_2) = \frac{S_2 x(t, \lambda_1, \lambda_2)}{S_1 x(t, \lambda_1)} = \frac{|a \star \psi_{\lambda_2}| \star \phi(t)}{a \star \phi(t)} + \widetilde{\epsilon}(t)$$

where $\widetilde{\epsilon}(t)$ 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 $\widetilde{\epsilon}$ produced by unvoiced sounds appears as small amplitude random fluctuations in Figure 7(c).

VII. Frequency Transposition Invariance

Audio signals within the same class may be transposed in frequency. Frequency transposition occurs when a single word is pronounced by different speakers. It is a complex phenomenon which affects the pitch and the spectral envelope. The envelope is translated on a logarithmic frequency scale but also deformed. We thus need a representation which is invariant to frequency translation on a logarithmic scale, and which also is stable to frequency deformations. 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 log-frequency.

MFCCs are computed from the log-mel-frequency spectrogram $\log Mx(t, \lambda)$ by calculating a DCT along the mel-frequency index γ for a fixed t [37]. 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 "quefrency" parameter. In MFCCs, high-quefrency coefficients are often set to zero, which is equivalent to averaging $\log Mx(t, 2^\gamma)$ along γ, which provides some frequency transposition invariance. The more high-quefrency coefficients are set to zero, the bigger the averaging and hence the more transposition invariance obtained, but at the expense of losing potentially important information.

The loss of information due to averaging along γ can be recovered by computing wavelet coefficients along γ. We thus replace the DCT by 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:

$$W^{\mathrm{fr}} z = \Big(z \star \phi^{\mathrm{fr}}(\gamma) \,,\, z \star \psi_q(\gamma) \Big)_{\gamma, q} \,. \quad (49)$$

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

Although the scattering transform along γ can be computed at any order, we restrict ourself to zero and first order

scattering coefficients, because it seems to be sufficient for classification. A first-order scattering transform of $z(\gamma)$ is calculated from

$$U^{\mathrm{fr}} z = \left(z(\gamma) \, , \, |z \star \psi_{q_1}(\gamma)| \right), \tag{50}$$

by averaging these coefficients along γ with ϕ^{fr}:

$$S^{\mathrm{fr}} z = \left(z \star \phi^{\mathrm{fr}}(\gamma) \, , \, |z \star \psi_{q_1}| \star \phi^{\mathrm{fr}}(\gamma) \right). \tag{51}$$

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. It has the same properties if we replace the time t by the log-frequency variable γ. Numerical experiments are implemented with Morlet wavelets ψ_{q_1} with $Q_1 = 1$.

Similarly to MFCCs, we apply a logarithm to normalized scattering coefficients so that multiplicative components become additive and can be separated by linear operators. This was shown to improve classification performance. The logarithm of a second-order normalized time scattering, at a frequency $\lambda_1 = 2^\gamma$ and a time t is

$$\log \widetilde{S} x(t, \gamma) = \begin{pmatrix} \log \widetilde{S}_1 x(t, 2^\gamma) \\ \log \widetilde{S}_2 x(t, 2^\gamma, \lambda_2) \end{pmatrix}_{\lambda_2} \tag{52}$$

This is a vector of signals $z(\gamma)$, where z depends on t and λ_2. Let us transform each $z(\gamma)$ by the frequency scattering operators U^{fr} or S^{fr}, defined in (50) and (51). Let $U^{\mathrm{fr}} \log \widetilde{S} x(t, \gamma)$ and $S^{\mathrm{fr}} \log \widetilde{S} x(t, \gamma)$ stand for the concatenation of these transformed signals for all t and λ_2. The representation $S^{\mathrm{fr}} \log \widetilde{S} x$ is calculated by cascading a scattering in time and a scattering in log-frequency. It is thus locally translation invariant in time and in log-frequency, and stable to time and frequency deformations. The interval of time-shift invariance is defined by the size of the time averaging window ϕ, whereas its frequency-transposition invariance depends upon the width of the log-frequency averaging window ϕ^{fr}.

Frequency transposition invariance is useful for speaker independent speech recognition or for transposition independant melody recognition, but it removes important information for other tasks such as speaker identification. The frequency transposition invariance, implemented by the frequency averaging filter ϕ^{fr}, should thus be adapted to the classification task. Next section explains that this can be done by replacing $S^{\mathrm{fr}} \log \widetilde{S} x(t, \gamma)$ by $U^{\mathrm{fr}} \log \widetilde{S} x(t, \gamma)$, and by optimizing the linear averaging at the supervised classification stage.

VIII. 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 more sophisticated state-of-the-art representations. Section VIII-A explains how to automatically adapt invariance parameters, while Sections VIII-B and VIII-C present results for musical genre classification and phone identification, respectively.

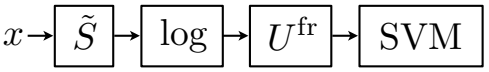

Fig. 8. *A time and frequency scattering representation is computed by applying a normalized temporal scattering $\widetilde{S}$ on the input signal $x(t)$, a logarithm, and a scattering along log-frequency without averaging.*

A. 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 normalized scattering transform $\widetilde{S} x$ of an input signal x is computed along time, over half-overlapping windows of size T. The log scattering vector for each time window is transformed along frequencies by the wavelet modulus operator U^{fr}, as explained in Section VII. 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 as input the vector of coefficients $\{U^{\mathrm{fr}} \log \widetilde{S} x(t, \gamma)\}_\gamma$, for each time frame indexed by 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 $\{U^{\mathrm{fr}} \log \widetilde{S} x(t, \gamma)\}_\gamma$, 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 [38].

The wavelet octave resolution Q_1 can also be adjusted at the supervised classification stage, by computing the time scattering for several values of Q_1 and concatenating all 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. The linear combination optimized by the SVM is a feature selection algorithm, which can select the best coefficients to discriminate any two classes. 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 time averaging window size T, for each signal class. For example, a phone duration may range from $10\mathrm{ms}$ to $200\mathrm{ms}$ and shorter phones are better discriminated with scattering coefficients calculated with smaller T. We thus concatenate

Representations	GTZAN	TIMIT
Δ-MFCC (T = 23ms)	19.3 $\pm$ 4.2	19.3
Δ-MFCC (T = 370ms)	17.8 $\pm$ 4.2	66.1
State of the art (excluding scattering)	9.4 $\pm$ 3.1 [8]	16.7 [42]
	$T = 370$ms	$T = 32$ms
Time Scat., $l = 1$	17.9 $\pm$ 4.2	18.5
Time Scat., $l = 2$	12.3 $\pm$ 2.7	17.7
Time Scat., $l = 3$	10.7 $\pm$ 2.0	18.7
Time & Freq. Scat., $l = 2$	10.3 $\pm$ 2.3	16.5
Adapt Q_1, Time & Freq. Scat., $l = 2$	9.0 $\pm$ 2.0	16.1
Adapt Q_1, T, Time & Freq. Scat., $l = 2$	8.1 $\pm$ 2.3	15.8

TABLE II
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 $T = 370$ms FOR GTZAN AND WITH $T = 32$ms FOR TIMIT.

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.

B. Musical Genre Classification

Scattering feature vectors are first applied to musical genre classification problem on the GTZAN dataset [39]. 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.

Preliminary experiments have demonstrated the efficiency of the scattering transform for music classification [40] and for environmental sounds [41]. These results are improved by letting the supervised classifier adjust the transform parameters to the signal classes. A set of feature vectors is computed over half-overlapping 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 II summarizes results with one run of ten-fold cross-validation. It gives the average error and its standard deviation.

Scattering classification results are first compared with results obtained with MFCC feature vectors. 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 $T = 23$ms, the Δ-MFCC error is 19.3%, which is reduced to 17.8% by increasing T to 370ms. Further increasing T does not reduce the error. State-of-the-art algorithms provide refined feature vectors to improve classification. Combining MFCCs with stabilized modulation spectra and performing linear discriminant analysis, [8] obtains an error of 9.4%, the best non-scattering result so far. A deep belief network trained on spectrograms [17], achieves 15.7% error with an SVM classifier. A sparse representation on a constant-Q transform [29], gives 16.6% error with an SVM.

Table II gives classification errors for different scattering feature vectors. For $l = 1$, they are composed of first-order time scattering coefficients computed using Morlet wavelets with $Q_1 = 8$ and $T = 370$ms. These vectors are similar to 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 $l = 2$, 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 $l = 3$, including these coefficients reduces the error marginally to 10.7%, at a significant computational and memory cost. We thus restrict ourselves to $l = 2$.

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 only 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 $T = 370$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 [8].

Replacing the SVM with more sophisticated classifiers can improve results. A sparse representation classifier applied to second-order time scattering coefficients reduces the error rate from 12.3% to 8.8%, as shown in [43]. Let us mention that the GTZAN database suffers from some significant statistical issues [44], which probably does not make it appropriate to evaluate further algorithmic refinements.

C. Phone Segment Recognition

The same scattering representation is tested for phone segment recognition with the TIMIT corpus [32]. The dataset contains 6300 phrases, each annotated with the identities, locations, and durations of its constituent phones. This task is much easier than continuous speech recognition, but provides an evaluation of scattering feature vectors for representing phone segments. Given the location and duration of a phone segment, the goal is to determine its class according to the standard protocol [45], [46]. 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 [47]. 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. In many cases, hidden Markov models or fixed time dilations

are applied to match different MFCC sequences, to account for the time-warping of the phone segment [45], [46]. Table II shows that $T = 23\text{ms}$ yields a 19.3% error which is much less than the 66.1% error for $T = 370\text{ms}$. Indeed, many phones have a short duration with highly transient structures and are not well-represented by wide time windows.

A lower error of 17.1% is obtained by replacing the SVM with a sparse representation classifier on MFCC-like spectral features [48]. Combining MFCCs of different window sizes and using a committee-based hierarchical discriminative classifier, [42] 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% [13].

Rows 4 through 6 of Table II gives the classification results obtained by replacing MFCC vectors with a time scattering transform computed with first-order Morlet 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 $l = 1$, 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 $l = 2$, 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 I. For the same reason, third-order coefficients provide even less information compared to the GTZAN case, and do not improve results.

For $l = 2$, cascading a log-frequency transposition invariance computed with a first-order frequency scattering transform of Section VII 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 and thus can suffer large-scale time-warping. This is done by aggregating scattering coefficients computed for $T = 32\text{ms}\,,\ 64\text{ms}\,,\ 128\text{ms}$. No explicit time warping is needed in the model. Thanks to the scattering deformation stability, supervised linear classifiers can indeed compute time-warping invariants which remain sufficiently informative.

IX. 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. Over windows of $T \approx 200\text{ms}$, signals recovered from first- and second-order scattering coefficients have a good audio quality. Normalized scattering coefficients characterizes amplitude modulations, and are stable to time-warping deformations. 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 segment identification on TIMIT.

Appendix

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 (44) that

$$|x \star \psi_{\lambda_1}(t)| \approx |\widehat{h}(\lambda_1)|\,|e \star \psi_{\lambda_1}(t)|\,a(t)\ . \tag{53}$$

Let us decompose

$$|e \star \psi_{\lambda_1}(t)| = \mathbb{E}(|e \star \psi_{\lambda_1}|) + \epsilon(t)\ , \tag{54}$$

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 wavelet with quadrature phase, 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 (54) and this equation in (53) shows that

$$|x \star \psi_{\lambda_1}(t)| \approx |\widehat{h}(\lambda_1)|\Big(\pi^{1/2}2^{-1}\|\psi_{\lambda_1}\|a(t) + a(t)\,\epsilon(t)\Big)\ . \tag{55}$$

When averaging with ϕ, we get

$$S_1 x(t,\lambda_1) \approx |\widehat{h}(\lambda_1)|\,\Big(\pi^{1/2}2^{-1}\|\psi_{\lambda_1}\|a \star \phi(t) + (a\,\epsilon) \star \phi(t)\Big)\ . \tag{56}$$

Suppose that $a(t)$ is not sparse, in the sense that

$$\frac{|a|^2 \star \phi(t)}{|a \star \phi|^2(t)} \sim 1\ . \tag{57}$$

It means that ratios between local $\mathbf{L}^2$ and $\mathbf{L}^1$ norms of a is of the order of 1. We are going to show that if $T^{-1} \ll \lambda_1 Q_1^{-1}$ then

$$\frac{\mathbb{E}(|(a\,\epsilon) \star \phi(t)|^2)}{\|\psi_{\lambda_1}\|^2\,|a \star \phi(t)|^2} \ll 1 \tag{58}$$

which implies (48). We give the main arguments to compute the order of magnitudes of the stochastic terms, but it is not a rigorous proof. For a detailed argument, see [49].

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 $h(t)$*

$$\mathbb{E}(|(za) \star h(t)|^2) \leq \sup_{\omega} \widehat{R}_z(\omega)\,|a|^2 \star |h|^2(t)\ . \tag{59}$$

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 [49] that $\sup_\omega \widehat{R}_\epsilon(\omega) \leq C(1-\pi/4)$, where $C \approx 1$. 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 (57) holds then

$$\frac{|a|^2 \star |\phi|^2(t)}{|a \star \phi(t)|^2} \sim \frac{1}{T} \tag{60}$$

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 (60), if $T^{-1} \ll \lambda_1 Q_1^{-1}$ it proves (58) and hence

$$S_1x(t,\lambda_1) \approx \frac{\pi^{1/2}}{2}\|\psi\|\, {\lambda_1}^{1/2}\, |\widehat{h}(\lambda_1)|\, a \star \phi(t)\ . \tag{61}$$

If $a \star \phi(t)$ is non-negligible then the scattering normalization (33) yields (48).

Let us now compute $S_2x(t,\lambda_1,\lambda_2) = ||x\star\psi_{\lambda_1}|\star\psi_{\lambda_2}|\star\phi(t)$. If $T^{-1} \ll \lambda_1 Q_1^{-1}$ then (61) together with (55) shows that

$$\frac{S_2x(t,\lambda_1,\lambda_2)}{S_1x(t,\lambda_1)} \approx \frac{|a\star\psi_{\lambda_2}|\star\phi(t)}{a\star\phi(t)} + \widetilde{\epsilon}(t)\,, \tag{62}$$

where

$$0 \leq \widetilde{\epsilon}(t) \leq \frac{2|(a\epsilon)\star\psi_{\lambda_2}|\star\phi(t)}{\pi^{1/2}\|\psi_{\lambda_1}\|\, a\star\phi(t)}\ . \tag{63}$$

Observe that

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

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

$$\mathbb{E}(|(a\epsilon)\star\psi_{\lambda_2}(t)|^2) \leq C(1-\pi/4)\, |a|^2 \star |\psi_{\lambda_2}|^2(t)\ . \tag{64}$$

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 (60), if (57) holds over time intervals of size Q_2/λ_2, then

$$\frac{|a|^2 \star |\psi_{\lambda_2}|^2(t)}{|a\star|\psi_{\lambda_2}||^2} \sim \frac{\lambda_2}{Q_2}. \tag{65}$$

Since $\|\psi_{\lambda_1}\|^2 \sim \lambda_1 Q_1^{-1}$ and $|\psi_{\lambda_2}| \star \phi(t) \sim \phi(t)$ when $Q_2/\lambda_2 \leq T$, it results from (63,64,65) that $0 \leq \mathbb{E}(\widetilde{\epsilon}(t)) \leq C\,(4/\pi-1)^{1/2}(\lambda_2\, Q_1)^{1/2}(\lambda_1\, Q_2)^{-1/2}$ with $C \sim 1$.

REFERENCES

[1] 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.

[2] H. Hermansky, "The modulation spectrum in the automatic recognition of speech," in *Proc. IEEE ASRU*, 1997, pp. 140–147.

[3] M. S. Vinton and L. E. Atlas, "Scalable and progressive audio codec," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 5. IEEE, 2001, pp. 3277–3280.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. K. Thompson and L. E. Atlas, "A non-uniform modulation transform for audio coding with increased time resolution," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 5. IEEE, 2003, pp. V–397.

[11] S. Mallat, "Group invariant scattering," *Commun. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.

[12] Y. LeCun, K. Kavukvuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE ISCAS*, 2010.

[13] H. Lee, P. Pham, Y. Largman, , and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. NIPS*, 2009.

[14] 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.

[15] 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.

[16] 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.

[17] P. Hamel and D. Eck, "Learning features from music audio with deep belief networks," in *Proc. ISMIR*, 2010.

[18] E. Battenberg and D. Wessel, "Analyzing drum patterns using conditional deep belief networks," in *Proc. ISMIR*, 2012.

[19] 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.

[20] 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.

[21] 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.

[22] J. Bruna and S. Mallat, "Invariant scattering convolution network," *IEEE Trans. Pattern Anal. Mach. Intell.*, to appear, http://arxiv.org/abs/1203.1513.

[23] L. Sifre and S. Mallat, "Rotation, scaling and deformation invariant scattering for texture discrimination," in *Proc. CVPR*, 2013.

[24] S. Mallat, *A wavelet tour of signal processing*. Academic Press, 1999.

[25] S. Schimmel and L. Atlas, "Coherent envelope detection for modulation filtering of speech," in *Proc. of ICASSP*, vol. 1, 2005, pp. 221–224.

[26] R. Turner and M. Sahani, "Probabilistic amplitude and frequency demodulation," in *Advances in Neural Information Processing Systems*, 2011, pp. 981–989.

[27] G. Sell and M. Slaney, "Solving demodulation as an optimization problem," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2051–2066, 2010.
[28] I. Waldspurger and S. Mallat, "Recovering the phase of a complex wavelet transform," CMAP, Ecole Polytechnique, Tech. Rep., 2012.
[29] M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. LeCun, "Unsupervised learning of sparse features for scalable audio classification," in *Proc. ISMIR*, 2011.
[30] J. Nam, J. Herrera, M. Slaney, and J. Smith, "Learning sparse feature representations for music annotation and retrieval," in *Proc. ISMIR*, 2012.
[31] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.
[32] 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.
[33] L. Lucy, "An iterative technique for the rectification of observed distributions," *Astron. J.*, vol. 79, p. 745, 1974.
[34] E. J. Candès, Y. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *arXiv preprint arXiv:1109.0573*, 2011.
[35] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," CMAP, Ecole Polytechnique, Tech. Rep., 2012.
[36] 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.
[37] 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.
[38] 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.
[39] 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.
[40] J. Andén and S. Mallat, "Multiscale scattering for audio classification," in *Proc. ISMIR*, Miami, Florida, Unites States, Oct. 24-28 2011, pp. 657–662.
[41] C. Baugé, M. Lagrange, J. Andén, and S. Mallat, "Representing environmental sounds using the separable scattering transform," in *Proc. IEEE ICASSP*, 2013.
[42] 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.
[43] X. Chen and P. J. Ramadge, "Music genre classification using multiscale scattering and sparse representations," in *Proc. CISS*, 2013.
[44] B. L. Sturm, "An analysis of the gtzan music genre dataset," in *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*. ACM, 2012, pp. 7–12.
[45] 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.
[46] 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.
[47] A. K. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, 1998.
[48] 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.
[49] J. Andén, "Time and frequency scattering representations," Ph.D. dissertation, Ecole Polytechnique, 2014.