AUDIO TEXTURE SYNTHESIS WITH SCATTERING MOMENTS

Joan Bruna

New York University
Courant Institute
New York, NY

Stéphane Mallat

Ecole Noramle Supérieure
Computer Science Department
Paris, France

ABSTRACT

We introduce an audio texture synthesis algorithm based on scattering moments. A scattering transform is computed by iteratively decomposing a signal with complex wavelet filter banks and computing their amplitude envelop. Scattering moments provide general representations of stationary processes computed as expected values of scattering coefficients. They are estimated with low variance estimators from single realizations. Audio signals having prescribed scattering moments are synthesized with a gradient descent algorithms. Audio synthesis examples show that scattering representation provide good synthesis of audio textures with much fewer coefficients than the state of the art.

***Index Terms*—** Audio synthesis, scattering moments, wavelets, texture.

1. INTRODUCTION

The representation of a non-Gaussian stationary process remains a fundamental issue of probability and statistics. Signal processing faces many such issues, in particular for auditory and image textures, which can be modeled as realizations of highly non-Gaussian processes. A random vector $X \in \mathbb{R}^N$ can be represented by a vector of generalized moments $\Phi X = \{\mathbb{E}(\phi_n(X)\}_n$ which project the distribution of X over multiple functions $\phi_n(x)$ with $x \in \mathbb{R}^N$. Random signal synthesis can then be performed by sampling the maximum entropy distribution, which is a Boltzmann distribution whose generalized moments are specified by ΦX. For most signal processing applications, one needs to estimate $\mathbb{E}(\phi_n(X)$ from a single realization of X, by replacing the expected value with a spatial or time average. We concentrate on on audio texture synthesis, which is an important application. The information loss of the representation can be checked by evaluating the perceptual quality of synthesized signals.

Second order moments lose essential perceptual information in audio and image signals because they provide no information on non-Gaussian behavior. High order moments are rarely used because their estimation from a single realization has a variance which is too large. Representations based on generalized moments have been proposed to represent and synthesize audio and image textures, often based on histograms of non-linear transformations of the signal [1, 2]. Simoncelli and McDermott have obtained particularly efficient results from covariance measurements at the output of multistage filter banks [3]. In the following we propose an audio texture representation and a synthesis algorithm based on scattering moments.

Scattering transforms have recently been introduced [4, 5, 6, 7] to represent audio signals and images, while providing state of the art results for texture discrimination, and genre recognition in audio [6]. A scattering transform iterates on complex wavelet transforms and modulus operators which compute their envelop. It has close relations with psychophysical and physiological models [8, 9, 10]. For stationary processes, it estimates a vector of expected values called scattering moments. This paper shows that scattering moments provide a compact representation of stationary processes, which encodes important non-Gaussian properties arising from multiscale amplitude and frequency modulations. This is demonstrated through audio synthesis.

Section 2 reviews the properties of scattering moments for auditory signals. An efficient audio synthesis algorithm is described in Section 3. Section 4 gives synthesis results on natural audio textures. Computations can be reproduced with a software available at *www.di.ens.fr/data/software/scatnet*.

Notations: $\widehat{x}(\omega) = \int x(t)\exp(-i\omega t)dt$ is the Fourier transform of $x(t)$. We denote $\mathbb{E}(X)$ the expected value of a stationary process $X(t)$ at any t, and $\sigma^2(X) = E(|X|^2) - E(X)^2$.

2. SCATTERING MOMENTS

A scattering transform characterizes transient structures through high order coefficients which capture modulation properties. They are computed by iterating on filter banks of complex wavelet filters.

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

2.1. Wavelet Filter Bank

A wavelet $\psi(t)$ is a band-pass filter. We consider a complex wavelet with a quadrature phase, whose Fourier transform satisfies $\widehat{\psi}(\omega) \approx 0$ for $\omega < 0$. We assume that the center frequency of $\widehat{\psi}$ is 1 and that its bandwidth is of the order of Q^{-1}. Wavelet filters centered at the frequencies $\lambda = 2^{j/Q}$ are computed by dilating ψ:

$$\psi_\lambda(t) = \lambda\, \psi(\lambda\, t) \;\text{ and hence }\; \widehat{\psi}_\lambda(\omega) = \widehat{\psi}(\lambda^{-1}\omega)\ . \quad (1)$$

We denote by Λ the index set of $\lambda = 2^{j/Q}$ over the signal frequency support, and we impose that these filters fully cover the positive frequencies

$$\forall \omega > 0\ ,\ 1 - \epsilon \leq \frac{1}{2} \sum_{\lambda \in \Lambda} |\widehat{\psi}_\lambda(\omega)|^2 \leq 1\ . \quad (2)$$

for some $\epsilon < 1$. The wavelet transform of a random process $X(t)$ is

$$WX = \{X \star \psi_\lambda(t)\}_{\lambda \in \Lambda}\ .$$

One can derive from (2) that the variance satisfies

$$\sigma^2(X)(1-\epsilon) \leq \sum_{\lambda \in \Lambda} \mathbb{E}(|X \star \psi_\lambda|^2) \leq \sigma^2(X)\ . \quad (3)$$

2.2. Scattering Moments

Scattering moments provide a representation of stationary processes, with expected values of a non-linear operator, calculated by iterating over wavelet transforms and a modulus. First order scattering coefficients are first order moments of wavelet coefficient amplitudes:

$$\forall la_1 \in \Lambda\ \ ,\ \ SX(\lambda) = \mathbb{E}(|X \star \psi_\lambda|)\ .$$

The Q-factor Q_1 adjusts the frequency resolution of these wavelets. First order scattering moments provide no information on the time-variation of the scalogram $|X \star \psi_{\lambda_1}(t)|$. It averages all audio modulations and transient events, and thus lose perceptually important information.

Second order scattering moments recover information on audio-modulations and transients by computing the wavelet coefficients of each $|X \star \psi_{\lambda_1}|$, and their first order moment:

$$\forall \lambda_2\ \ ,\ \ SX(\lambda_1, \lambda_2) = \mathbb{E}(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|)\ .$$

These multiscale variations of each envelop $|X \star \psi_{j_1}|$, specify the amplitude modulations of $X(t)$ [6]. The second family of wavelets ψ_{j_2} typically have a Q-factor $Q_2 = 1$ to accurately measure the sharp transitions of amplitude modulations. Scattering coefficients have a negligible amplitude for $\lambda_2 > \lambda_1$ because $|X \star \psi_{\lambda_1}|$ is then a regular envelop whose frequency support is below λ_2. Scattering coefficients are thus computed only for $\lambda_2 < \lambda_1$.

Applying more wavelet transform envelops defines scattering moments at any order $m \geq 1$:

$$\overline{S}X(\lambda_1, ..., \lambda_m) = \mathbb{E}(|\ |X \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}|)\ . \quad (4)$$

By iterating on the inequality (3), one can verify [4] that the Euclidean norm of scattering moments

$$\|\overline{S}X\|^2 = \sum_{m=1}^{\infty} \sum_{(\lambda_1,...,\lambda_m) \in \Lambda_m} |\overline{S}X(\lambda_1, ..., \lambda_m)|^2. \quad (5)$$

satisfies

$$\|\overline{S}X\|^2 \leq \sigma^2(X)\ .$$

Expected scattering coefficients are first moments of non-linear functions X and thus depend upon high order moments of X [4]. But as opposed to high order moments, the scattering representation is computed with wavelet transforms and modulus operators, which do not amplify the variability of X. It results into low-variance estimators.

Scattering moments are estimated by replacing the expectation with a time averaging over the signal support. Suppose that $X(t)$ is defined for $0 \leq t < N$. With periodic border extensions, we compute empirical averages

$$\widehat{S}X(\lambda_1, ..., \lambda_m) = N^{-1} \sum_{t=1}^{N} |\ |X \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}(t)|\ . \quad (6)$$

For most audio textures, the energy of the scattering vector $\|\overline{S}X\|^2$ is concentrated over first and second order moments [6]. We thus only compute $\widehat{S}X(\lambda_1)$ and $\widehat{S}X(\lambda_1, \lambda_2)$ for $1 \leq \lambda_1 = 2^{j_1/Q_1} \leq N$ and $1 \leq \lambda_2 = 2^{j_2/Q_2} < \lambda_1$. Scattering moments estimators have large variance at the lowest frequencies because the wavelet coefficient amplitudes are highly correlated in time. These higher variance estimators are removed by keeping only the frequencies λ_1 and λ_2 above a fixed frequency N_0. We thus compute $Q_1 \log_2(N/N_0)$ first order scattering moments and $Q_1 Q_2 (\log_2 N/N_0)^2/2$ second order scattering moments.

Scattering transforms have been extended along the frequency variables to capture frequency variability and provide transposition invariant representations [6]. Transpositions refer to translations along a log frequency variable. For audio synthesis, this frequency transformation will only be performed on first order coefficients. We denote $\gamma = \log_2 \lambda_1$, and define wavelets $\bar{\psi}_{\bar{\lambda}}(\gamma)$ having an octave bandwidth of $Q = 1$. The corresponding wavelet transform is thus computed with convolutions along the log-frequency variable γ.

The scalogram is now considered as a function of γ for each fixed time t:

$$F_t(\gamma) = |X \star \psi_{2^\gamma}(t)|\ .$$

Second order frequency scattering moments are the first order moments of the wavelet coefficients of $F_t(\gamma)$ computed along γ:

$$\overline{S}X(\lambda_1, \bar{\lambda}_2) = \mathbb{E}(|F_t \star \bar{\psi}_{\bar{\lambda}_2}(\log_2 \lambda_1)|)\ ,$$

This expected value is estimated with a time averaging

$$\widehat{S}X(\lambda_1, \bar{\lambda}_2) = N^{-1} \sum_{t=1}^{N} |F_t \star \bar{\psi}_{\bar{\lambda}_2}(\log_2 \lambda_1)| \, . \quad (7)$$

If $K = Q_1 \log_2(N/N_0)$ is the total number of first order scattering moments, the number of second order frequency scattering coefficients is αK, where α is an oversampling constant which is set to 2 in our experiments.

3. SCATTERING SYNTHESIS

We present a gradient descent algorithm on the scattering domain to adjust scattering moments estimated from available observations.

A maximum entropy distribution satisfying a set of moment conditions is a Gibbs distribution defined by the Boltzmann theorem. Sampling this distribution is possible with the Metropolis-Hastings algorithms but it is computationally very expansive in high dimension. This algorithm is often approximated with a gradient descent algorithm. It is initialized with a Gaussian white noise realization, whose moments are progressively adjusted by the gradient descent [11, 1, 3].

Let $Y(t)$ be the realization of an auditory texture of N samples. A vector of first order and second order scattering moment estimators $\widehat{S}X$ is computed with (6). This vector may also include second order frequency scattering moments (7). To synthesize a new audio signal X such that $\widehat{S}X = \widehat{S}Y$, we start with a realization of white Gaussian noise X_0. At each iteration n, we want to minimize

$$E(X) = \frac{1}{2} \|\widehat{S}X_n - \widehat{S}Y\|^2. \quad (8)$$

A gradient descent computes

$$X_{n+1} = X_n - \gamma \nabla E(X_n) = X_n - \gamma \partial \widehat{S}X_n^T (\widehat{S}X_n - \widehat{S}Y) \, , \quad (9)$$

where $\partial \widehat{S}X_n$ is the Jacobian of $\widehat{S}X$ with respect to X, evaluated at X_n, and γ is a gradient step, which is kept fixed at a sufficiently small value for the sake of simplicity.

The minimization of (8) is a non-linear least squares problem. The Levenberg-Marquardt Algorithm (LMA) [12] significantly accelerates the convergence. It replaces $\partial \widehat{S}X_n^T$ in (9) by the pseudoinverse

$$\partial \widehat{S}X_n^{\dagger} = (\partial \widehat{S}X_n^T \partial \widehat{S}X_n)^{-1} \partial \widehat{S}X_n^T ,$$

which requires computing a pseudoinverse on each iteration. The LMA typically requires 20 iterations to reach a relative approximation error of 10^{-2} and 40 to reach 10^{-4}, tested on the collection of auditory textures described in next section.

4. NUMERICAL EXPERIMENTS

The audio scattering synthesis algorithm is tested on a dataset of natural sound textures of McDermott and Simoncelli, available at [13]. It is a collection of 15 sound textures, of 7 seconds each, sampled at 20 KHz, thus including $N \sim 10^5$ samples. Our synthesis results are available at [14].

McDermott and Simoncelli [3] have constructed an audio representation based on physiological models of audition. Similarly to a scattering transform, it uses two constant-Q filter banks. The first set of *cochlea* filters consists in 30 complex bandpass filters. Their envelop is first compressed with a contractive nonlinearity and then redecomposed with a new filter bank. They extract a collection of 1500 coefficients, comprising marginal moments of each cochlea envelop and their corresponding modulation bands, as well as pairwise cross-correlations across different cochlea and modulation bands. In [15], the authors used a similar model to produce a texture representation with about 800 coefficients.

Scattering audio synthesis is performed with much fewer coefficients. With $Q_1 = 4$ and $N_0 = 2^2$ there are $Q_1 \log_2 N/N_0 = 46$ first order moments, $Q_1 Q_2 (\log_2 N/N_0)^2/2 = 266$ second order moments and $2 \cdot 46 = 92$ frequency scattering moments The total representation thus has 402 coefficients. Figure 1 shows the scalogram of signals recovered from first order moments only or first and second order moments. Reconstructions from first order moments are essentially realizations of Gaussian processes. They do not capture the transient and impulsive structures of the textures, such as the hammer or the applause. When second order scattering moments are included, the reconstructed textures contain these highly non-Gaussian phenomena, which produce highly realistic synthesized sounds. Scattering moments have the ability to capture processes with irregular spectra, such as the jackhammer, as well as wideband phenomena such as fire cracking or applause.

Figure 2 shows that frequency scattering moments correlate and thus synchronize the amplitude variations across frequency bands. This is necessary to accurately reproduce transient structures in textures. The synthesis of wide-band textures can be further improved by combining scattering moments computed with dyadic wavelets having $Q_1 = 1$. It adds 120 coefficients which further constraint the frequency interferences created by time varying modulations.

5. CONCLUSIONS

A texture audio synthesis is performed with a gradient descent algorithm which progressively adjusts the scattering moments of a signal. Good perceptual reconstructions are obtained with fewer coefficients than state of the art algorithms.

First and second order scattering moments are thus efficient texture descriptors; on the one hand, they are sufficiently informative so that realizations with similar coefficients have good perceptual similarity. On the other hand, they are consistent: realizations of the same process (hence perceptually similar) have similar scattering representations, as opposed to high order moments.

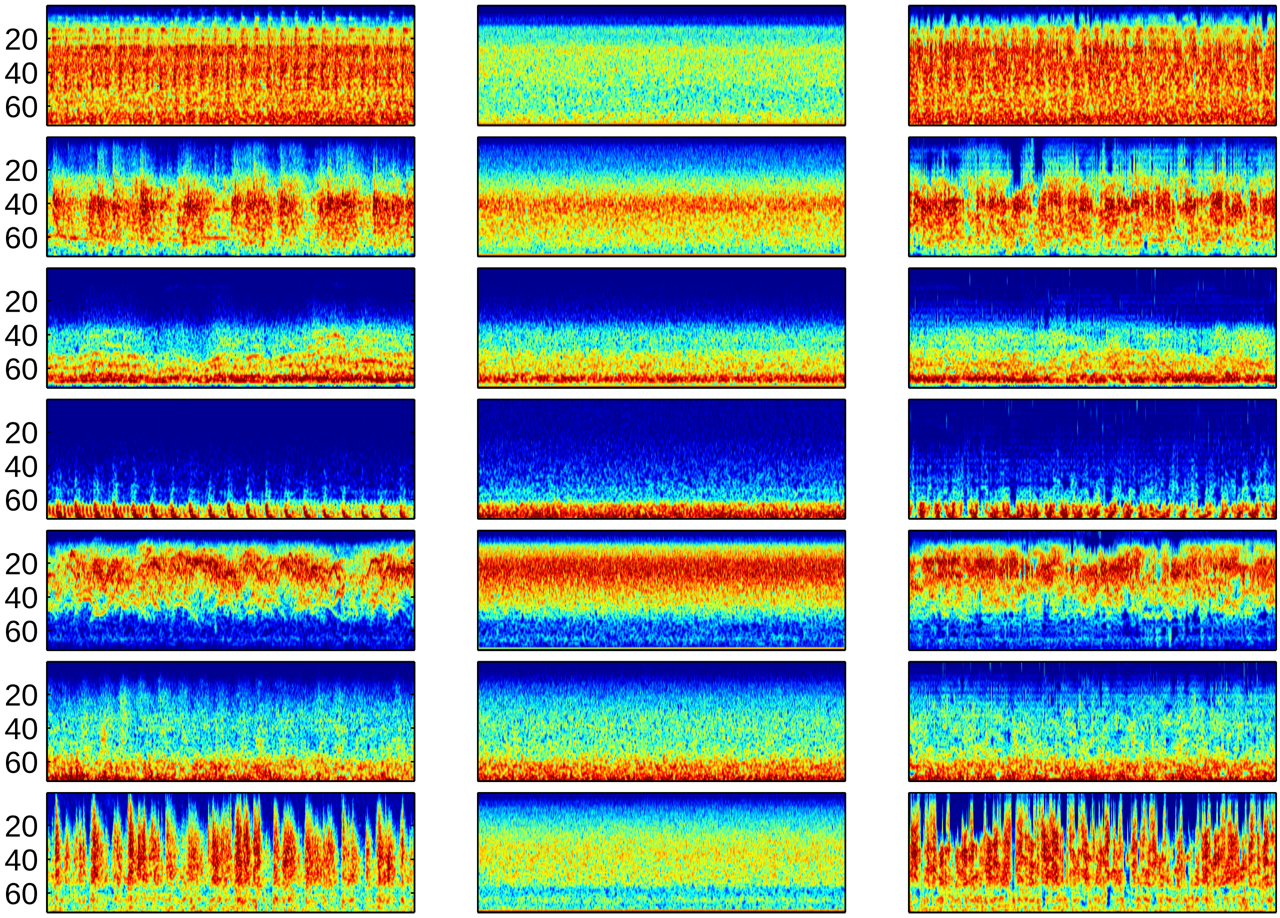

Fig. 1. Each image the scalogram of an audio recording: time along the horizontal axis and log-frequency up to 10KHz along the vertical axis. Left column: original audio textures from [13]. Middle column: Reconstruction from 1st order time scattering moments Right column: reconstruction from 1st and 2nd order time scattering moments. The sounds are produced (from to to bottom) by jackhammer, applause, wind, helicopter, sparrows, train, rusting paper.

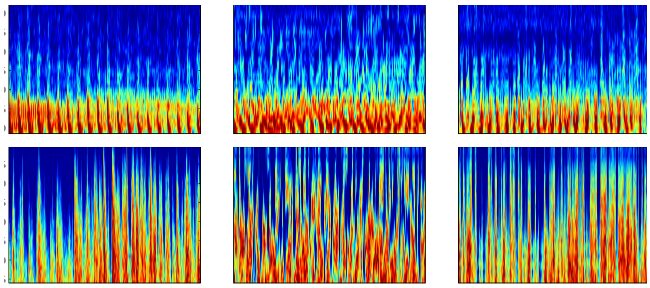

Fig. 2. Impact of frequency scattering moments. Left column: original signals. Middle column: synthesis from first and second order time scattering moments. Right column: synthesis obtained by adding frequency scattering moments. Observe how without frequency scattering, the subbands tend to decorrelate, which prevents synthesizing impulsive phenomena. The sounds are produced by a helicopter and rusting paper. More examples available at *cims.nyu.edu/∼bruna*.

6. REFERENCES

[1] J. Portilla and E. Simoncelli, "A Parametric Texture Model based on Joint Statistics of Complex Wavelet Coefficients," *IJCV*, 2000.

[2] P. Huber and B. Kleiner, "Statistical Methods for investigating phase relations in stochastic processes," *IEEE Trans on Audio and Electroacoustics*, 1976.

[3] J. McDermott and E. Simoncelli, "Sound Texture Perception via statistics of the auditory periphery: Evidence from Sound Synthesis," *Neuron*, 2011.

[4] S. Mallat, "Group Invariant Scattering," *Communications in Pure and Applied Mathematics*, 2012.

[5] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE transactions of PAMI*, 2012.

[6] J. Anden and S. Mallat, "Deep scattering spectrum," *IEEE transactions of Signal Processing*, 2013.

[7] L. Sifre and S. Mallat, "Combined Scattering for Rotation Invariant Texture Analysis," *CVPR*, 2013.

[8] 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.

[9] 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.

[10] M. Slaney and R. Lyon, *Visual representations of speech signals*, chapter On the importance of time–a temporal representation of sound, pp. 95–116, M. Cooke, S. Beet and M. Crawford (Eds.) John Wiley and Sons, 1993.

[11] S. Zhu, Y. Wu, and D. Mumford, "Minimax Entropy Principle and Its Application to Texture Modeling," *Neural Computation*, 1997.

[12] D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, 1963.

[13] J. McDermott, "http://www.cns.nyu.edu/∼jhm/," .

[14] J. Bruna, "http://cims.nyu.edu/∼bruna/," .

[15] R. McWalter and T. Dau, "Analysis of the auditory system via Sound Texture Synthesis," *International Conference on Acoustics*, 2013.