WAVELET SCATTERING ON THE SHEPARD PITCH SPIRAL

*Vincent Lostanlen, Stéphane Mallat**

Dept. of Computer Science,
École normale supérieure
Paris, France
vincent.lostanlen@ens.fr

ABSTRACT

1. INTRODUCTION

Spectrogram-based pattern recognition algorithms, such as sparse coding [Abdallah Plumbley 2005] and Nonnegative Matrix Factorization [Smaragdis Brown 2003], are widespread in audio signal processing. They are designed to approximate their input by a linear combination of few data-driven templates. Musical chords, for example, are expected to get decomposed into individual notes.

However, most natural sounds cannot be factorized as amplitude-modulated fixed spectra: notably, continuous changes in pitch (e.g. vibrato, glissando) as well as in spectral envelope (e.g. attack transients, formantic transitions) have a joint time-frequency structure that cannot be matched to a single spectral atom. Time-varying, under-constrained generalizations have been devised to address this shortcoming [Hennequin et al. 2011], but their high number of parameters prevents their robustness in challenging polyphonic contexts.

Instead of specifying probabilistic priors to help the convergence [Fuentes et al. 2013], we aim to design a template-free, nonlinear, mid-level representation, that natively disentangles the time variabilities of pitch and spectral envelope.

The central idea to our representation is that the former correspond to rigid motions along the log-frequency axis, whereas the latter affect the relative amplitude of harmonics across neighboring octaves. This distinction can be conceptually emphasized by arranging the log-frequency axis in a spiral, hence aligning frequency bins that share the same musical pitch class or "chroma" [Shepard 1964]. By means of a multivariable wavelet transform (see Fig. 1), which consists of joint time-chroma-octave convolutions, changes in pitch and spectral envelope are respectively captured as angular and radial motions on the spiral.

The contributions of this paper are:

- the introduction of the Shepard spiral scattering transform as a cascade of wavelet operators,
- a nonstationary formulation of the source-filter convolutional model relying on time warps, and its factorization in the wavelet scalogram,
- an approximate closed-form expression of Shepard spiral scattering coefficients, showing that variabilities in pitch and spectral envelope get jointly linearized, and stably appear as energy maxima.
- a visualization of these coefficients in Berio's *Sequenza V*, revealing extended instrumental techniques.

* This work is supported by the ERC InvariantClass 320959.

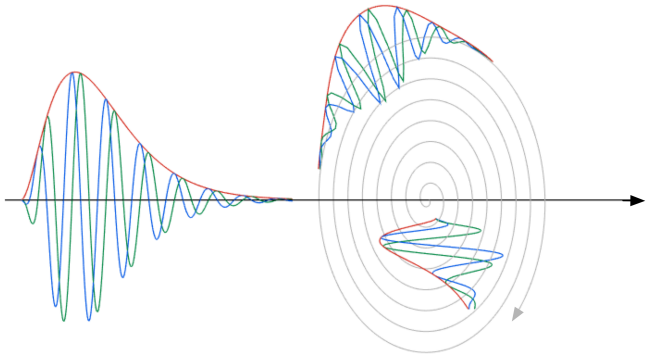

Figure 1

2. SHEPARD SPIRAL SCATTERING

Let $\psi(t) = |\psi|(t)\mathrm{e}^{2\pi i t}$ a "mother wavelet" of dimensionless center frequency 1 and bandwidth Q^{-1}. The quality factor Q is an integer in the typical range 12–24. Center frequencies of the subsequent wavelet filter bank are of the form $\lambda_1 = 2^{j_1 + \frac{\chi_1}{Q}}$, where the indices $j_1 \in \mathbb{Z}$ and $\chi_1 \in \{1 \dots Q\}$ respectively denote octave and chroma. The Fourier transform $\widehat{\psi}(\omega)$ of $\psi(t)$ is dilated by resolutions λ_1 to obtain wavelets $\widehat{\psi_{\lambda_1}}$ in the frequency domain:

$$\widehat{\psi_{\lambda_1}}(\omega) = \widehat{\psi}(\lambda^{-1}\omega) \quad \text{i.e.} \quad \psi_{\lambda_1}(t) = \lambda_1 \psi(\lambda_1 t).$$

The wavelet transform of an audio signal $x(t)$ is defined as the array of convolutions $x * \psi_{\lambda_1}(t)$ for every audible frequency λ_1. The modulus of the resulting signals, called *scalogram*, localize the power spectrum of $x(t)$ around the log-frequencies $\log_2 \lambda_1 = j_1 + \frac{\chi_1}{Q}$ over durations $2Q\lambda_1^{-1}$, trading frequency resolution for time resolution:

$$x_1(t, \log_2 \lambda_1) = |x * \psi_{\lambda_1}|\,(t).$$

The constant-Q transform (CQT) $S_1 x$ corresponds to a low-pass filtering of x_1 with a window $\phi(t)$ of size T.

$$S_1 x(t, \log_2 \lambda_1) = x_1 * \phi_T(t) = |x * \psi_{\lambda_1}| * \phi_T(t).$$

There is a well-known dilemma in choosing T. Too small, the constant-Q matrix lacks invariance to time shifts, which will

prevent any learning step to generalize from $S_1 x$; too large, discriminative information is discarded.

In order to combine the best of both worlds, the scattering transform recovers finer time scales than T with a second filterbank of wavelets $\psi_{\lambda_2}(t)$ of center frequencies λ_2, and applies complex modulus to improve regularity [Andén Mallat 2011].

$$x_2(t, \log_2 \lambda_1, \log_2 \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| (t)$$

Also known as *amplitude modulation spectrum* [Thompson Atlas 2003], the three-way array x_2 is then averaged in time to achieve as much invariance as the constant-Q spectrum $S_1 x$:

$$S_2 x(t, \log_2 \lambda_1, \log_2 \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| (t) * \phi_T(t).$$

The concatenated scattering representation $Sx = \{S_1 x, S_2 x\}$ has proven to achieve higher accuracy in music genre classification as well as phoneme recognition [Andén et Malla 2011] than audio features derived from $S_1 x$ only, such as Mel-frequency cepstral coefficients (MFCC).

It should be noted that, while the definition above successfully describes the average spectral envelope and amplitude modulation of a signal, it decomposes and averages each frequency band separately. Hence, it cannot properly characterize the joint time-frequency structure of natural sounds.

Cepstral coefficients, which pictures pitch on a line, does not involve pitch circularity. Conversely, chroma features are insensible to octave transpositions.

$$\log_2 \lambda_1 = \lfloor \log_2 \lambda_1 \rfloor + \{\log_2 \lambda_1\}$$

$$\Psi_{\lambda_2}(t, \log_2 \lambda_1) = \psi_\alpha(t) \times \psi_\beta(\log_2 \lambda_1) \times \psi_\gamma(\lfloor \log_2 \lambda_1 \rfloor)$$

The integer part $\lfloor \log_2 \lambda_1 \rfloor$ is the octave index (related to perceived pitch height), whereas the fractional part $\{\log_2 \lambda_1\}$ is related to pitch chroma.

3. DEFORMATIONS OF THE SOURCE-FILTER MODEL

4. CONCLUSIONS

Future work will be devoted to evaluating the discriminative power of Shepard spiral scattering coefficients over a variety of classification pipelines. Our representation also encompass automatic music transcription, perceptual similarity learning, and new audio transformations as potential applications.