TEMPLATES FOR DAFX-15, TRONDHEIM, NORWAY

*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 "chroma", i.e. musical pitch class [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.

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 $\chi_1 \in \{1 \ldots Q\}$ respectively denote octave and chroma.

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

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 scattering transform has been introduced in signal classification to achieve discriminative invariants to time shifts, while ensuring stability to small time warps. Invariance up to some maximal time shift T is achieved with a lowpass filter ϕ_T of time support T. The outcome of this filtering are the constant-Q transform (CQT) coefficients $S_1 x$, indexed by time t and log-frequency $\log_2 \lambda_1$:

$$S_1 x(t, \log_2 \lambda_1) = x_1 * \phi_T(t).$$

This averaging discards all variability in x_1 at finer time scales than T, that is, higher modulation frequencies than $T^{-}1$. To recover them, the scattering transform convolves x_1 with a second filterbank of wavelets $\psi_{\lambda_2}(t)$ of center frequencies λ_2, and applies complex modulus to improve regularity:

$$x_2(t, \log_2 \lambda_1, \log_2 \lambda_2) = |x_1 * \psi_{\lambda_2}|\,(t)$$

Also known as *amplitude modulation spectrum* three-way array x_2
$S_2 x(t, \log_2 \lambda_1, \log_2 \lambda_2) = x_2 * \phi_T(t)$

* This work was supported by the XYZ Foundation