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

We present a new reprensetation of sounds that liearizes the dynamics of pitch chroma and pitch height, while remaining stable to deformations in the time-frequency plane. It is an instance of the scattering transform, a generic operator which cascades wavelet convolutions and modulus nonlinearities. It is derived from the Shepard pitch spiral, in that convolutions are performed in time, log-frequency (correlated to pitch chroma) and octave index (correlated to pitch height).

1. INTRODUCTION

Spectrogram-based pattern recognition algorithms, such as sparse coding [1] and Nonnegative Matrix Factorization [2], 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 [3], but their high number of parameters prevents their robustness in challenging polyphonic contexts.

Instead of specifying probabilistic priors to help the convergence [4], 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" [5]. 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,

* This work is supported by the ERC InvariantClass 320959.

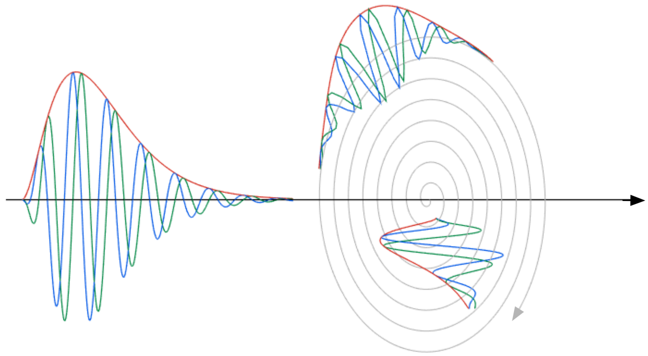

Figure 1

- 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

2.1. Time 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. 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:

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

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 filter-bank of wavelets $\psi_{\lambda_2}(t)$ of center frequencies λ_2, and applies complex modulus to improve regularity [6]. The wavelets $\psi_{\lambda_2}(t)$ have a quality factor in the range 1–2, though we choose to keep the same notation ψ for simplicity.

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

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

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

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 [6] than audio features derived from $S_1 x$ only, such as Mel-frequency cepstral coefficients (MFCC).

2.2. Joint time-frequency scattering

Due to the constant-Q property, Sx is stable to small time warps of $x(t)$, as long as they do not exceed Q^{-1}, i.e. one semitone. This implies that small modulations, such as tremolo and vibrato, are accurately linearized in rate and depth [7].

However, the definition above is unstable to the variability in pitch and spectral envelope, for which the activation of frequency bands is highly correlated in time. To stabilize x_2 with respect to these variations, Andén [8] has redefined the wavelets ψ_{λ_2}'s as two-dimensional functions of both time and log-frequency, indexed by pairs $\lambda_2 = (\alpha, \beta)$, where α is measured in Hertz and β is measured in cycles per octaves.

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

The equation below introduces a "joint time-frequency scattering" transform, as opposed to the plain "time scattering" transform of Equation 2:

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

The joint time-frequency scattering transform corresponds to the "cortical transform" introduced by Shamma to formalize his findings in auditory neuroscience.

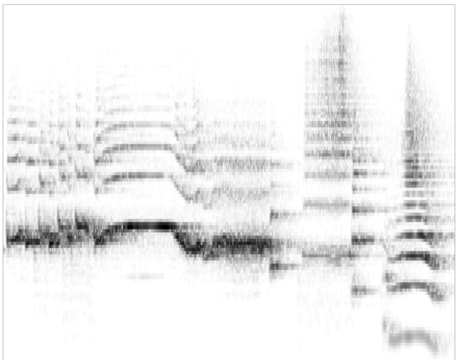

Figure 2

2.3. Spiral scattering

— DRAFT BELOW—

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

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

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

$$x(t) = [e_\theta * h_\zeta(t)]$$

4. CONCLUSIONS

The spiral model is well-known in music theory and experimental psychology. However, existing methods in audio signal processing do not fully take advantage from its richness, as they either picture pitch on a line (e.g. MFCC) or on a circle (e.g. chroma features).

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.

5. REFERENCES

[1] S. Abdallah and M. Plumbley, "Polyphonic music transcription by non-negative sparse coding of power spectra," in *Proc. ISMIR*, 2004, vol. 510, pp. 10–14.

[2] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, 2003.

[3] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model nonstationary audio events," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744–753, 2011.

[4] B. Fuentes, R. Badeau, and G. Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 9, pp. 1854–1866, 2013.

[5] R. Shepard, "Circularity in Judgments of Relative Pitch," 1964.

[6] J. Andén and S. Mallat, "Deep Scattering Spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[7] Joakim Andén and Stéphane Mallat, "Scattering representation of modulated sounds," *Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx-12)*, , no. 3, pp. 15–18, 2012.

[8] Joakim Andén, *Time and frequency scattering for audio classification*, Ph.D. thesis, École Polytechnique, 2014.