Wavelet scattering on the Shepard pitch spiral

Vincent Lostanlen

April 2015

1 Introduction

Many audio processing challenges rely on the characterization of highly transient phenomena up to time scales of about one second. Yet, classical representations of timbre, such as the Mel-frequency cepstral coefficients or the chroma features, lose most of their discriminability as soon as the window size T exceeds 50 ms. Consequently, it is necessary to build an aggregation strategy on top of the short-term features, before feeding them a classifier. On one hand, all-purpose clustering techniques have shown to perform quite poorly on time-frequency representations, due to the fast variability in natural sounds. On the other, deep learning architectures, which consist of an alternated cascade of convolutional units and pointwise nonlinearities, happen to be much more suited to modeling long-range interactions. Indeed, each layer in a feed-forward network is optimized to integrate locally the information dynamics in its input, while retaining as much disciminative information as possible.

Attractive as it may seem, this data-driven approach requires a large amount of labeled examples to converge to a satisfying solution. Moreover, the training step is computationally intensive and may not work at all if its hyperparameters are not carefully cross-validated over the training set. Last, but not least, deep networks are designed to restrict the support of the convolutional units ; whereas this constraint helps feature specialization and alleviates computations, it also prevents the grasp of long-range interactions within the frequential axis. For instance, harmonic sounds exhibit a comb-like spectrum of partials, whose positions and amplitudes are highly correlated. These sounds are ubiquitous in speech and music processing, and their timbral evolution is known to be essential to audio classification. Learned representations, which are often contrained to operate on small time-frequency patches, do not capitalize on that phenomenon. In this article, we introduce a new representation for audio classification that is able to capture the spectral envelope as well as the fine variations in pitch. To do so, we revisit the pitch spiral model formalized by cognitive scientist Roger Shepard to explain the circularity of pitch classes between octaves. After computing a multiscale spectrogram of the audio signal, we cascade three wavelet transforms along different variables: time, log-frequency, and octave index. We show that the two latter transforms detect angular and radial motion in the pitch spiral, hence characterizing the respective evolutions of pitch contour and formantic structure. The subsequent representation, called spiral scattering transform, is a two-layer nonlinear network of multidimensional convolutions. Unlike a neural network, it involves no preliminary training step, since its convolutional units are specified in closed form by the geometry of the Shepard pitch spiral. To illustrate its discriminative power, we perform audio synthesis from time-invariant coefficients by iterated backpropagation.

2