MULTIPLE-F0 ESTIMATION AND NOTE TRACKING FOR PIANO MUSIC FOR MIREX 2014 USING TEMPORAL EVOLUTION INFORMATION

Andrea Cogliati, Zhiyao Duan
AIR Lab, Department of Electrical and Computer Engineering
University of Rochester
andrea.cogliati@rochester.edu, zduan@ur.rochester.edu

ABSTRACT

In this submission for MIREX 2014 we utilize a novel piano music transcription algorithm based on the temporal evolution of piano notes. Most existing transcription algorithms, especially those based on Non-negative matrix factorization and Probabilistic latent component analysis, operate on a spectrogram on a frame-by-frame basis, i.e., they do not consider the temporal evolution of the notes of musical instruments, which in the case of percussive instruments is very characteristic. In our work, we propose a spectrogram factorization algorithm that uses the full spectrogram of sampled notes as template and a greedy algorithm to detect the templates to activate for a given chord. Combining this algorithm with a robust onset detection method will lead to a more accurate piano music transcription system. We also propose a new metric for spectrogram similarity.

1. INTRODUCTION

Automatic Music Transcription (AMT) is the process of extracting a music notation representation from an audio file. The output of AMT can be a full musical score, or an intermediate representation, like a MIDI piano roll, which includes note pitches, onsets, offsets and, possibly, the instrument playing the notes. The proposed method analyzes a monaural, polyphonic recording of a piano piece and extracts a MIDI piano roll.

Non-negative matrix factorization (NMF) is a method for factoring a large non-negative matrix (a matrix with non-negative elements) into the product of two, low-rank non-negative matrices

$$V \simeq W \times H. \tag{1}$$

The factorization is approximate. The two matrices W and H represent a dictionary of semantically meaningful elements (templates) and their point of activation.

NMF has been applied to source separation and AMT. The matrix V is the spectrogram of an audio file, the matrix W contains a set of spectral templates of different musical events (ideally the spectra of individual notes in the audio file) while H indicates which templates are activated in each frame of V.

One of the drawbacks of this approach is that it analyzes a spectrogram on a frame-by-frame basis, i.e., it does not consider the natural temporal evolution of the notes of musical instruments. Thus, each template is the averaged spectrogram of a single musical event and the reconstruction assumes that all the partials evolve in the same way. This is not true for piano notes; higher frequency partials decay faster than lower frequency ones (see fig. 1 and fig. 2).

Finally, since the algorithms for updating the factorizing matrices operate in the continuous domain, musically informed constraints like sparsity and polyphony can only be applied in the post-processing stage, which is not ideal.

2. PROPOSED METHOD

The proposed method operates in three stages:

1. Onset detection
2. Pitch identification
3. Post-processing.

The first stage analyzes the audio file looking for note onsets. The percussive nature of the piano helps detecting the onsets, even at softer dynamics. Onsets temporally close enough can be considered as a single onset without excessive loss of precision in the successive stage.

The second stage analyzes the audio between two successive onsets and identifies the pitches in the chord using a greedy algorithm.

The post-processing stage optimizes the results of the second stage comparing the pitches in successive chords. I.e., the same pitch identified in two successive chords can be a false positive due to reverb or sustain pedal.

2.1 Onset detection

The onset detection step uses a neural network to analyze the normalized spectral flux. The linear magnitude spectrogram from the original audio is generated using a STFT with 2048 samples, Hamming window and a hop size of 128 samples. The spectral flux is calculated by summing all the positive bin-to-bin differences in two successive frames.

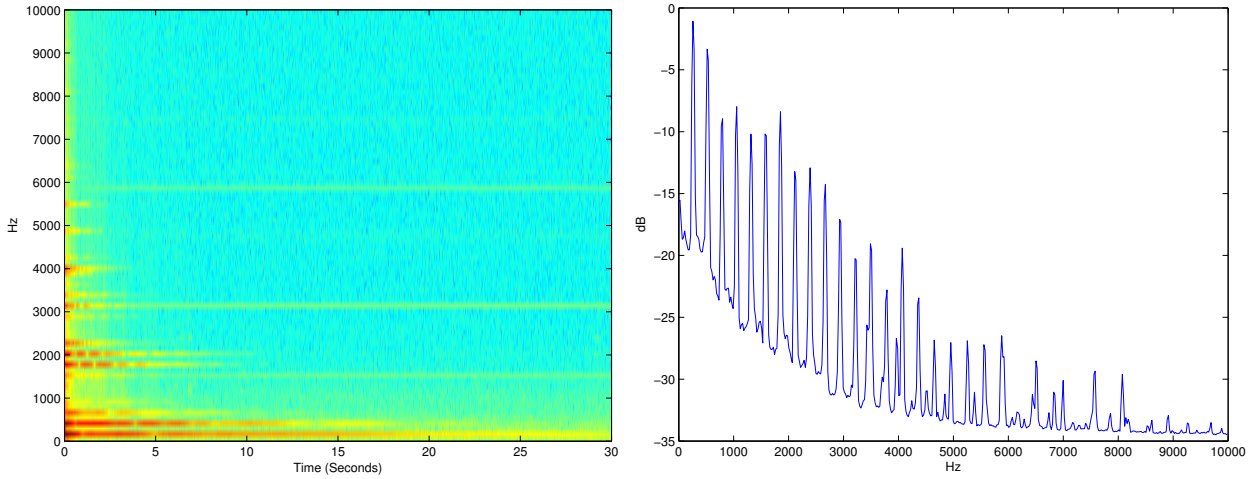

Figure 1. Spectrogram and long time average spectrum (LTAS) of a single piano note (C4).

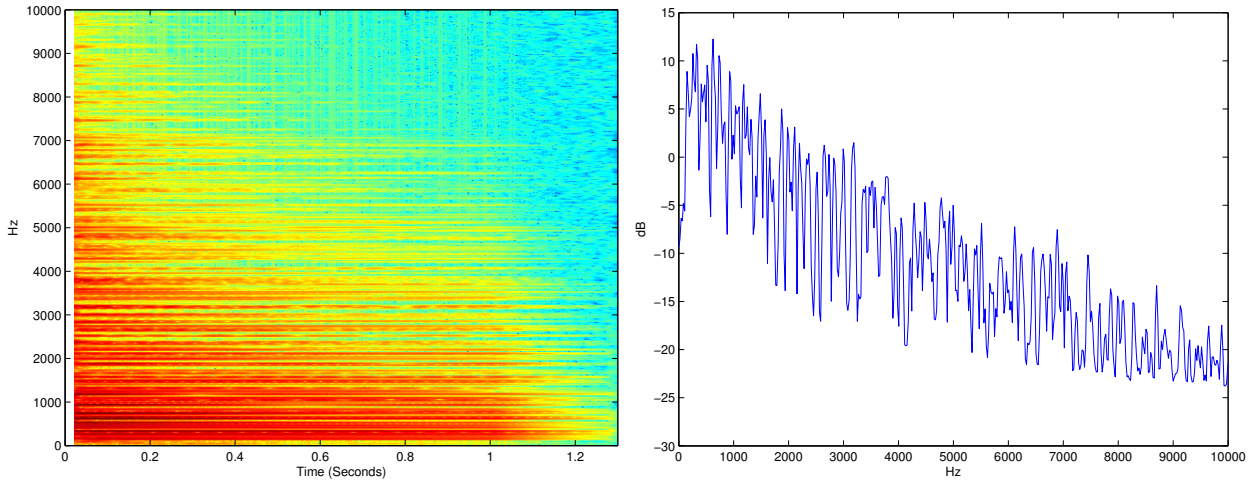

Figure 2. Spectrogram and LTAS of a piano chord.

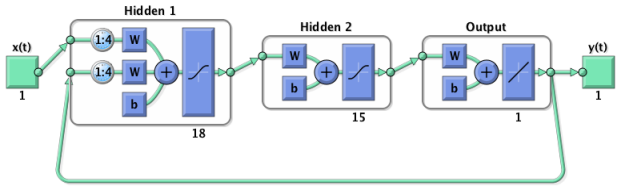

Figure 3. Artificial neural network used for onset detection.

The normalized spectral flux is then processed by a nonlinear autoregressive network with exogenous inputs (NARX) neural network with two hidden layers, with 18 and 15 neurons respectively, and 4 delays (see fig. 3). The neural network has been trained using 10 pieces from the MAPS database [2].

2.2 Pitch identification

The individual pitches in a chord are identified using a greedy algorithm. A dictionary of templates is generated from the University of Iowa Musical Instrument Samples [1], in which the individual notes of a piano have been recorded at three different dynamic levels. The template log-frequency, linear magnitude spectrograms are calculated using a constant Q transform with 36 bins per octave and a hop size of 1024 samples. The templates are limited to 128 columns. The resulting spectrograms are filtered to extract only the active partials. The linear magnitude of piano partials show a characteristic curve which can be approximated by a summation of two decaying exponentials. Using a curve fitting algorithm, only the bins which can be approximated by a summation of two decaying exponentials are preserved. All the other bins are set to 0.

A spectrogram with the same parameters is generated from each chord detected in the previous step. The spectrogram is filtered using the same approach used in the template generation.

The greedy algorithm compares each note in the dictionary and computes the cost function. The note with the lowest cost is selected. Then the algorithm tries to add a second note and, if the cost function decreases by at least a certain amount, the second note is selected. The algorithm stops when an additional note does not lower the cost func-

tion.

2.2.1 *Spectrogram similarity*

The spectrogram similarity is measured using a cost function based on a scaled L^2-norm. Given V, the spectrogram of the original audio, and R, the reconstructed spectrogram, the distance of the two spectrograms is measured by

$$\sum_{ij} \left(\frac{V_{ij} - R_{ij}}{\log_1 0(V_{ij} + 1) + 1} \right)^2. \quad (2)$$

2.3 Post-processing

The post-processing step is not currently implemented but it can be used to increase the precision by incorporating prior musical knowledge.

3. REFERENCES

[1] The University of Iowa Musical Instrument Samples [1]

[2] Emiya, V., Badeau, R. and David, B.: "Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 6, pp. 1643-1654, Aug. 2010.

[1] `http://theremin.music.uiowa.edu/MIS.html`