MIREX 2014 ENTRY: CONVEX NON-NEGATIVE MATRIX FACTORIZATION

Oriol Nieto
Music and Audio Research Lab
New York University
oriol@nyu.edu

Tristan Jehan
The Echo Nest
tristan@echonest.com

ABSTRACT

This extended abstract describes the structural segmentation submission to MIREX 2014 of an algorithm that uses Convex Non-negative Matrix Factorization (C-NMF) in order to both identify and label the large scale segment boundaries of a given audio track. The method employs harmonic and timbral features synchronized to a detected beat, and it trivially extracts the structural information from the activation matrix obtained with C-NMF when using the combined features as input. This implementation is open source and available for public download [1] .

1. INTRODUCTION

The MIR task of Music Structure Analysis aims at automatically identifying the large-scale, contiguous non-overlapping segments that compose a specific audio track. Once these segments are estimated, they are labeled based on their acoustic similarity. It is well-established in the literature that music segments can be classified under three different types: novelty, homogeneous, and repetitive [7]. The former are segments whose boundaries are determined based on sudden local variations in one of the audio features (e.g. Checkerboard-like kernel method [3]). The homogeneous segments have a relatively uniform audio quality from start to end (e.g. Shift Invariance Probabilistic Latent Component Analysis [8]). Finally, the repetitive segments re-occur across the piece, and this quality can be exploited in order to identify them (e.g. Fitness Measure [5]).

This year we submitted a music structure analysis algorithm to MIREX (**NJ1**) that is able to identify homogeneous segments, and then label them based on their resemblance. This method is based on [6], which adds a convex constraint to the Non-negative Matrix Factorization technique, resulting in more meaningful and stable cluster centroids that yield better music segmentation when using audio features as input to the matrix factorization process.

[1] https://github.com/urinieto/SegmenterMIREX2014

2. AUDIO FEATURES

The package Essentia [1] is used to compute the audio features from a one-channel audio file sampled at 44.1 kHz (these features are the same as the ones used for other algorithms submitted to MIREX by one of the authors, see submissions **NB1**, **NB2**, and **NB3**). More specifically, we use Harmonic Pitch Class Profile features —a harmonic representation that is intended to be less noisy than regular chromagrams—, and a Hann analysis window of size 2048 samples with a 50% overlap. Additionally, we also use Essentia to compute 13 Mel Frequency Cepstral Coefficients to capture the timbre of the track. We combine the HPCP with the MFCC by simply stacking them and producing a 25 dimensional feature vector. Furthermore, we use the Multi Feature Beat Tracker [9] implementation included in Essentia to estimate the beats, and the aggregation described in [2] to obtain the beat-synchronous representation. Finally, a median filter is applied to these synchronous features before factorizing them such that C-NMF can better capture the homogeneous segments.

3. C-NMF FOR MUSIC SEGMENTATION

The unsupervised machine learning method of Convex NMF was previously introduced under the music structure framework in [6]. Traditional NMF aims at factorizing a given matrix $X \in \mathbb{R}^{p \times N}$ into two matrices $F \in \mathbb{R}^{p \times r}$ and $G \in \mathbb{R}^{r \times N}$, such that $X \approx FG$, and r is the rank of decomposition. This classic approach has also been used for music segmentation, initially introduced in [4]. Alternatively, C-NMF adds a constraint to F such that its columns $\mathbf{f}$ become convex combinations of the observations of X:

$$\mathbf{f}_j = \mathbf{x}_1 w_{1j} + \cdots + \mathbf{x}_N w_{Nj} = X\mathbf{w}_j \quad j \in [1:r] \quad (1)$$

This results in $F = XW$, where $W \in \mathbb{R}^{N \times r}$, which makes the columns $\mathbf{f}_j$ interpretable as weighted cluster *centroids*, representing, in our case, better section prototypes of the musical piece (in F) and sharper activations (in G). Consequently, C-NMF can be formally described as: $X \approx XWG = FG$.

In our submission we modified the original algorithm by using the actual audio features as input X (as opposed to the self-similarity matrix of the features discussed in [6]). In Figure 1 an example is shown, where X represents the beat-synchronous HPCP of a given track using $r = 4$. As

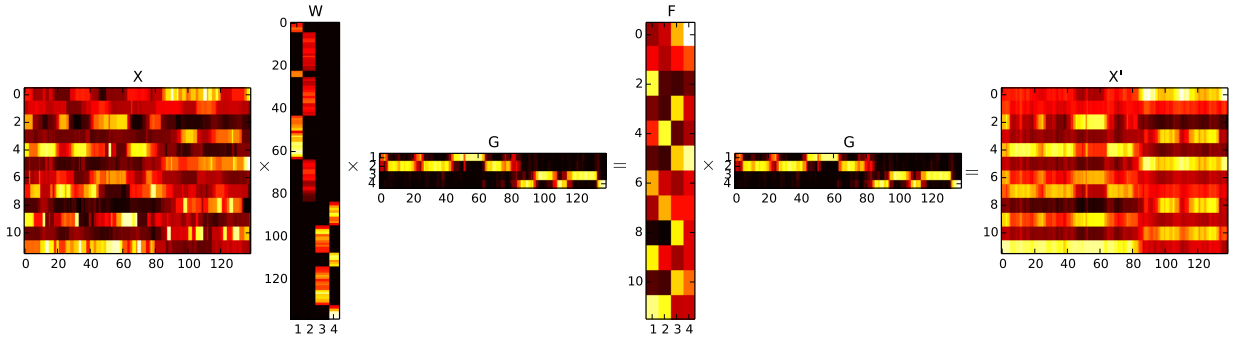

Figure 1. Visualization of the C-NMF process using the HPCP of the song *And I Love Her* by The Beatles.

it can be seen, the activations matrix G contains a meaningful interpretation of the structural segments of the track represented by X, already labeled based on their similarity. We can extract this information by thresholding G and applying a low pass filter to remove spurious segments.

Empirically, we found that a lower r was beneficial for the boundary identification stage, whereas a higher r yielded better results for the labeling subtask. Moreover, using HPCP plus MFCC for identifying the segments was also advantageous, while it was better to simply use HPCP for the labeling subtask. Therefore, we apply C-NMF twice for each track: one time to estimate the segment boundaries and a second time to group the segments in musically meaningful labels. For the detailed set of parameters we refer the reader to the open-source implementation of this algorithm, whose link is found in the abstract of this document.

4. REFERENCES

[1] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra. Essentia: An Audio Analysis Library for Music Information Retrieval. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, pages 493–498, Curitiba, Brazil, 2013.

[2] Daniel P. W. Ellis and Graham E. Poliner. Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking. In *Proc. of the 32nd IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1429–1432, Honolulu, HI, USA, 2007.

[3] Jonathan Foote. Automatic Audio Segmentation Using a Measure Of Audio Novelty. In *Proc. of the IEEE International Conference of Multimedia and Expo*, pages 452–455, New York City, NY, USA, 2000.

[4] Florian Kaiser and Thomas Sikora. Music Structure Discovery in Popular Music Using Non-Negative Matrix Factorization. In *Proc. of the 11th International Society of Music Information Retrieval*, pages 429–434, Utrecht, Netherlands, 2010.

[5] Meinard Müller and Frank Kurth. Towards Structural Analysis of Audio Recordings in the Presence of Musical Variations. *EURASIP Journal on Advances in Signal Processing*, 2007(1):089686, 2007.

[6] Oriol Nieto and Tristan Jehan. Convex Non-Negative Matrix Factorization For Automatic Music Structure Identification. In *Proc. of the 38th IEEE International Conference on Acoustics Speech and Signal Processing*, pages 236–240, Vancouver, Canada, 2013.

[7] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-Based Music Structure Analysis. In *Proc of the 11th International Society of Music Information Retrieval*, pages 625–636, Utrecht, Netherlands, 2010.

[8] Ron Weiss and Juan Pablo Bello. Unsupervised Discovery of Temporal Structure in Music. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1240–1251, 2011.

[9] José R. Zapata, Matthew E. P. Davies, and Emilia Gómez. Multi-feature Beat Tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):816–825, 2013.