TOWARDS MODELING AND DECOMPOSING LOOP-BASED ELECTRONIC MUSIC

Patricio López-Serrano **Christian Dittmar** **Jonathan Driedger** **Meinard Müller**
International Audio Laboratories Erlangen, Germany
patricio.lopez.serrano@audiolabs-erlangen.de

ABSTRACT

Electronic Music (EM) is a popular family of genres which has increasingly received attention as a research subject in the field of MIR. A fundamental structural unit in EM are loops—audio fragments whose length can span several seconds. The devices commonly used to produce EM, such as sequencers and digital audio workstations, impose a musical structure in which loops are repeatedly triggered and overlaid. This particular structure allows new perspectives on well-known MIR tasks. In this paper we first review a prototypical production technique for EM from which we derive a simplified model. We then use our model to illustrate approaches for the following task: given a set of loops that were used to produce a track, decompose the track by finding the points in time at which each loop was activated. To this end, we repurpose established MIR techniques such as fingerprinting and non-negative matrix factor deconvolution.

1. INTRODUCTION

With the advent of affordable electronic music production technology, various loop-based genres emerged: techno, house, drum'n'bass and some forms of hip hop; this family of genres is subsumed under the umbrella term *Electronic Music* (EM). EM has garnered mainstream attention within the past two decades and has recently become a popular subject in MIR: standard tasks have been applied to EM (structure analysis [17]); new tasks have been developed (breakbeat analysis and resequencing [7, 8]); and specialized datasets have been compiled [9].

A central characteristic of EM that has not been extensively considered is its sequencer-centric composition. As noted by Collins [4], *loops* are an essential element of EM: loops are short audio fragments that are "generally associated with a single instrumental sound" [3]. Figure 1 illustrates a simplified EM track structure similar to that encouraged by digital audio workstations (DAWs) such as *Ableton Live* [1]. The track starts with the activation of a drum loop (blue, bottom row). After one cycle, a melody loop (yellow, middle row) is added, while the drum loop continues to play. A third layer—the bass (red, top row)—is activated in the third cycle. Over the course of the track, these loops are activated and deactivated. An important observation is that all appearances of a loop are identical; a property that can be modeled and exploited in MIR tasks. In particular, we consider the task of *decomposing* an EM track: given the set of loops that were used to produce a track and the final, *downmixed* version of the track itself, we wish to retrieve the set of timepoints at which each loop was activated.

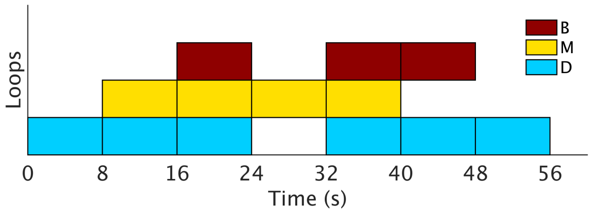

Figure 1. A condensed EM track built with three loop layers: drums (D), melody (M) and bass (B). Each block denotes the activation of the corresponding pattern during the time it spans.

This work offers three main contributions. First, we review the production process of EM and how it leads to the prototypical structure outlined previously (Section 2). Second, we propose a simplified formal model that captures these structural characteristics (Section 3). Third, we use our model to approach the EM decomposition task from two angles: first, we interpret it within a standard retrieval scenario by using fingerprinting and diagonal matching (Section 4). Our second approach is based on non-negative matrix factor deconvolution (NMFD), a technique commonly used for audio source separation (Section 5). We summarize our findings and discuss open issues in Section 6.

2. STRUCTURE AND PRODUCTION PROCESS

Unlike other genres, EM is often produced by starting with a single distinct musical pattern [19] (also called *loop*) and then adding and subtracting further musical material to shape the tension and listener's expectation. An EM track is built by combining layers (with potentially different lengths) in looping cyclical time—where the overall form corresponds to the multitrack layout of sequencers and digital audio workstations (DAWs) [4]. Figure 1 provides a simple example of such a track (total duration 56 s), consisting of three layers or loops: drums (D), bass (B) and melody (M), each with a duration of 8 s. We will be using this track as a running example to clarify the points made throughout this paper.

A common characteristic of EM tracks is their relative *sparseness* or low timbral complexity during the intro and outro—in other words, a single loop is active. This practice is rooted in two facts: Firstly, EM tracks are conceived not as isolated units, but rather as part of a seamless mix (performed by a DJ), where two or more tracks are overlaid together. Thus, in what could be termed *DJ-friendly* tracks [4], a single, clearly percussive element at the beginning and end facilitates the task of beat matching [3] and helps avoid unpleasantly dense transitions. We have constructed our running example following this principle: in Figure 1, the only active layer during the intro and outro is the drum loop (bottom row, blue).

The second reason for having a single-layer intro is that this section presents the track's main elements, making the listener aware of the sounds [3]. Once the listener has become familiar with the main musical idea expressed in the intro, more layers are progressively brought in to increase the tension (also known as a *buildup*), culminating in what Butler [3] designates as the track's *core*: the "thicker middle sections" where all loop layers are simultaneously active. This is reflected in Figure 1, where the melody is brought in at 8 s and the bass at 16 s, overlapping with uninterrupted drums. After the core has been reached, the majority of layers are muted or removed—once again, to create musical anticipation—in a section usually known as *break* or *breakdown* (see the region between 24–32 s in Figure 1, where only the melody is active). To release the musical tension, previous loops are reintroduced after the *breakdown*, (seconds 32–40, Figure 1) only to be gradually removed again, arriving at the outro. In the following sections we will develop a model that captures these structural characteristics and provides a foundation for analyzing EM tracks.

3. SIMPLIFIED MODEL FOR EM

In Section 2 we illustrated the typical form of loop-based electronic music. With this in mind, our goal is to analyze an EM track's structure. More specifically, our method takes as input the set of loops or patterns that were used to produce a track, as well as the final, *downmixed* version of the track itself. From these, we wish to retrieve the set of timepoints at which each loop was activated within the track. We begin by formalizing the necessary input elements.

Let $V \in \mathbb{R}^{K \times M}$ be the feature representation of an EM track, where $K \in \mathbb{N}$ is the feature dimensionality and $M \in \mathbb{N}$ represents the number of elements or frames along the time axis. We assume that the track was constructed from a set of R patterns $P^r \in \mathbb{R}^{K \times T^r}$, $r \in [0 : R-1] := \{0, \ldots, R-1\}$. The parameter $T^r \in \mathbb{N}$ is the number of feature frames or observations for pattern P^r. In practice, the patterns can have different lengths—however, without loss of generality, we define their lengths to be the same $T := T^0 = \ldots = T^{R-1}$, which could be achieved by adequately zero-padding shorter patterns until they reach the length of the longest. Based on this assumption, the patterns can be grouped into a pattern tensor $P \in \mathbb{R}^{K \times R \times T}$. In the case of our running example, seen in Figure 1, $T \mathrel{\hat{=}} 8$ s and the number of patterns is $R = 3$. Consequently, the subdimension of the tensor which refers to a specific pattern with index r is $P^r := P(\cdot,\ r,\ \cdot)$ (i. e., the feature matrix for either (D), (M), or (B) in our example); whereas $P_t := P(\cdot,\ \cdot,\ t)$ refers to frame index t simultaneously in all patterns.

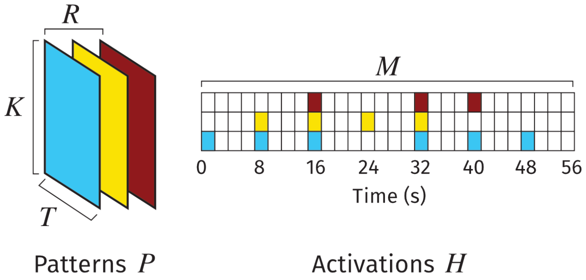

Figure 2. **(Left)**: Tensor P with three patterns (drums, bass, melody). **(Right)**: Activation matrix H with three rows; the colored cells denote an activation of the corresponding pattern.

In order to construct the feature representation V from the pattern tensor P, we require an activation matrix $H \in \mathbb{B}^{R \times M}$ with $\mathbb{B} := \{0, 1\}$, such that

$$V \mathrel{\hat{=}} \sum_{t=0}^{T-1} P_t \cdot \overset{t\rightarrow}{H}, \tag{1}$$

where $\overset{t\rightarrow}{(\cdot)}$ denotes a frame shift operator [18]. Figure 2 depicts P and H as constructed for our running example. The model assumes that the sum of pattern signals and their respective transformations to a feature representation are linear, which may not always be the case. The additive assumption of Eq. 1 implies that no time-varying and/or non-linear effects were added to the mixture (such as compression, distiortion, or filtering), which are often present in real-world EM. Aside from this, we specify a number of further constraints below.

The devices used to produce early EM imposed a series of technical constraints which we formalize here. Although many of these constraints were subsequently eliminated in more modern equipment and DAWs, they have been ingrained into the music's aesthetic and remain in use up to the present day.

Non-overlap constraint: A pattern is never superimposed with itself, i. e., the distance between two activations of any given pattern is always equal to or greater than the pattern's length. Patterns are loaded into a device's mem-

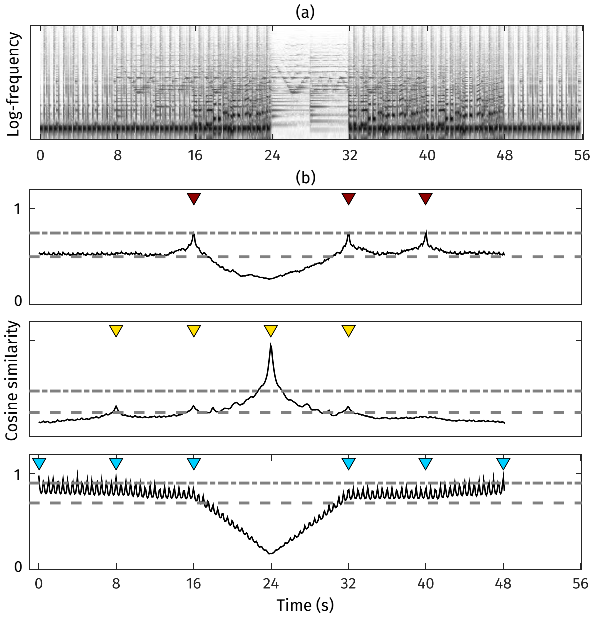

Figure 3. **(a)**: Log-frequency spectrogram for entire track. **(b)**: Matching curves computed using cosine similarity for drums, melody, and bass (bottom to top). The dashed line represents the curve's global mean; the dash-dotted line is the GT mean (see Section 4.4 for definition of *gain*). Colored triangles indicate GT loop activation positions.

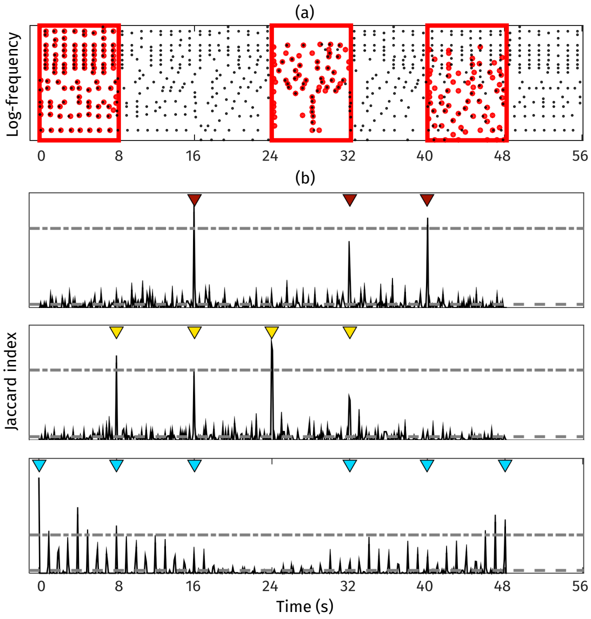

Figure 4. **(a)**: Log-frequency spectral peak map for the entire track (black dots) and for each query (red dots enclosed in red, from left to right: drums, melody, and bass). **(b)**: Matching curves computed with the Jaccard index and each pattern as a query for drums, melody, and bass (bottom to top).

ory and triggered by a sequencer—usually without further activation signals before it has played through to the end. If a pattern P^r is activated at time $m \in [0 : M-1]$, then $H^r(m) \neq 0 \Rightarrow H^r(m+1) = \ldots = H^r(m+T-1) = 0$.
Length constraint: As noted by [4], multiple layers in EM are complementary, creating aggregate effects and capable of being independently inserted and removed. For this reason, we make the simplifying assumption that $T := T^0 = T^1 = \ldots = T^{R-1}$, i. e., that all patterns have the same length.
T-grid constraint: Motivated by the use of centralized MIDI clocks and the fixed amount of musical time available on prevalent devices such as drum machines (which typically allow programming one musical measure at a time, in 16 steps), we enforce a timing grid which restricts the possible activation points in H. In Figure 1, patterns are always introduced and removed at multiples of 8 s.
Amplitude constraint: We assume that a pattern is always activated with the same *intensity* throughout a track, and therefore each row r in the activation matrix H fulfills $H^r := H(r, \cdot) \in \mathbb{B}^{1\times M}$.

4. FINGERPRINT-BASED EM DECOMPOSITION

In the running example, multiple patterns are overlaid in different configurations to form the track. If we know *a priori* which patterns are included and wish to find their respective activation positions, we need a technique capable of identifying an audio query within a database where further musical material is superimposed. We first examine log-frequency spectrograms and diagonal matching as a baseline approach, and continue with audio fingerprinting techniques based on spectral peaks in combination with various similarity measures. In Section 5 we discuss an alternative approach based on NMFD. The running example is constructed with one audio file for each pattern and a generic EM track arrangement seen in Figure 1. The complete track is generated in the time domain by summing the individual patterns that are active at a given point in time. All audio files have been downmixed to mono with a sampling rate $F_s = 22050$ Hz.

4.1 Diagonal Matching

We implement the diagonal matching procedure outlined in [13, pp. 376–378] to measure the *similarity* between each query pattern P^r and the track feature matrix V. In simple terms, to test if and where the query $P^r = (P^r_0, \ldots, P^r_{T-1})$ is contained in $V = (V_0, \ldots, V_{M-1})$, we shift the sequence P^r over the sequence V and locally compare P^r with suitable subsequences of V. In general, let $\mathcal{F}$ be the feature space (for example, $\mathcal{F} = \mathbb{R}^K$ in the case of log-frequency spectrograms). A similarity measure $s : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R} \cap [0, 1]$ between two feature frames will yield a value of 1 if the query is identical to a certain region of the database, and 0 if there is no resemblance at all.

4.2 Baseline Procedure

For each individual pattern, as well as the entire track, we compute an STFT $\mathcal{X}$ with the following parameters: block size $N = 4096$, hop size $H = N/2$ and a Hann window. From these, magnitude spectrograms (abbreviated as MS) are computed and mapped to a logarithmically spaced frequency axis with a lower cutoff frequency of 32 Hz, an upper cutoff frequency of 8000 Hz and spectral selectivity of 36 bins per octave (abbreviated LS). Under these STFT settings, the 36-bin spectral selectivity does not hold in the two lowest octaves; however, their spectral peak contribution is negligible. Preliminary experiments have shown that this musically meaningful feature representation is beneficial for the matching procedure both in terms of efficiency and accuracy. We begin with a baseline experiment (Figure 3), using LS and cosine similarity:

$$\textit{Cosine}: s_{\cos}(u,v) := 1 - \frac{\langle u|v\rangle}{||u|| \cdot ||v||},\ u,v \in \mathbb{R}^K \setminus \{0\}. \tag{2}$$

Notice that the clearest peak is produced by the melody activation at 24 s (Figure 3b, middle row), which occurs without any other patterns being overlaid. The three remaining activation points for the melody have a very low gain relative to their neighboring values. The matching curve for the drums (Figure 3b, bottom row) displays a coarse downwards trend starting at 0 s and reaching a global minimum at 24 s (the point at which the drum pattern is not activated in our example); this trend is reversed as the drums are added again at 32 s. The internal repetitivity (or self-similarity) of the drum pattern causes the periodic peaks seen throughout the matching curve. Overall, it is evident from all three curves that the combination of LS with cosine similarity is insufficient to capture the activations when multiple patterns are superimposed—motivating our next experimental configuration which uses spectral peak maps.

4.3 Fingerprinting with Peak Maps

Although our scenario is slightly different to that of audio fingerprinting and identification, both require a feature representation which captures an individual pattern's characteristics despite the superposition of further sound sources. To this end, we use spectral peak maps as described in [13]. Conceptually, we are following an early approach for loop retrieval inside hip hop recordings which was presented in [20] and later refined in [2]. Their method is based on a modification of the fingerprinting procedure originally described in [21].

For each time-frequency bin in the respective LS, a rectangular analysis window is constructed. The maximum value within each window is kept (with the value 1 on the output) and all neighbors are set to 0 on the output. In Figure 4a we show the spectral peak map for the entire track (black dots) and the query peak map for each query pattern (red dots in red rectangles). These log-frequency peak maps populate a pattern tensor $P \in \mathbb{B}^{K \times R \times T}$, where $K = 286$. Thus P^r corresponds to the peak map for the r^{th} pattern, while V corresponds to the entire track.

	Gain		*Pearson*	
	μ	σ	μ	σ
MS/cos	1.72	0.31	0.13	0.05
LS/cos	1.57	0.29	0.11	0.05
PLS/cos	19.46	10.45	0.52	0.18
PLS/inc	21.69	11.90	0.51	0.19
PLS/Jac	19.54	9.76	0.53	0.18

Table 1. Results for diagonal matching experiments with magnitude spectrograms (MS), log-frequency spectrograms (LS), and log-frequency peak maps (PLS) using the cosine, inclusion and Jaccard similarity measures. Each column shows the mean and variance for peak gain and Pearson correlation.

In addition to the cosine measure defined in Eq. 2, we test different similarity measures s:

$$\textit{Jaccard}: s_{\text{Jac}}(u,v) := 1 - \frac{||u \wedge v||}{||u \vee v||},\ u,v \in \mathbb{B}^K, \tag{3}$$

$$\textit{Inclusion}: s_{\text{inc}}(u,v) := 1 - \frac{||u \wedge v||}{||u||},\ u,v \in \mathbb{B}^K, \tag{4}$$

where we set $\frac{0}{0} := 1$. The inclusion metric aims to quantify the extent to which the query is contained or *included* in the database and has a similar definition to the Jaccard index.

4.4 Evaluation

We use two measures to quantify how well the matching curves capture the pattern activations. For the first measure, termed *gain*, we compute the average of the activation values at the ground truth (GT) activation points: in Figures 3b and 4b, these locations are marked by colored triangles, corresponding to each loop in the running example; their mean value is shown as a dash-dotted line. We also compute the mean value for the entire curve (dashed line) and use the ratio between these two means in order to assess the quality of the matching curve. Ideally, the curve assumes large values at the GT activation points and small values elsewhere, resulting in a larger gain. As a second measure we take the *Pearson correlation* between a computed matching curve and its corresponding row H^r in the GT activation matrix, where the activation points have a value of 1, and 0 elsewhere. Again, a high Pearson correlation reflects high matching curve quality.

We generated a set of patterns used to build prototypical EM tracks. To foster reproducible research, we produced them ourselves, avoiding potential copyright issues—they are available under a Creative Commons Attribution-ShareAlike 4.0 International license and can be obtained at the companion website [1]. We chose seven prominent EM subgenres such as *big beat*, *garage* and *drum'n'bass* (in a tempo range between 120–160 BPM). For each subgenre, we generated four patterns in the categories of drums, bass, melody and additional effects.

[1] `https://www.audiolabs-erlangen.de/resources/MIR/2016-ISMIR-EMLoop`

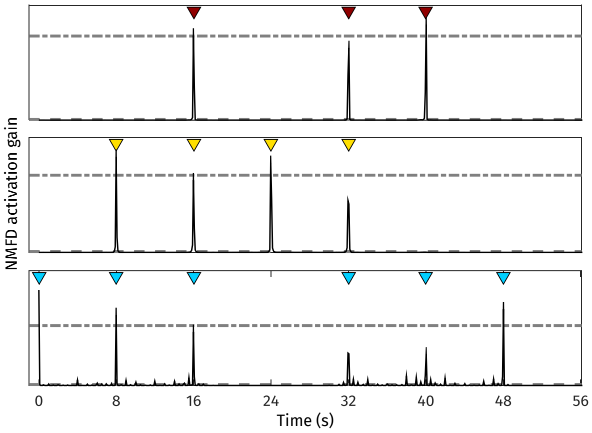

Figure 5. Activation curves learned by NMFD applied to the magnitude spectrogram of the running example. The dashed lines represent the mean value for the complete curve, but are very close to the x-axis.

As stated by Wang [21], spectral peak maps are robust to superposition of multiple sources; a fact which becomes clear when comparing Figures 3b and 4b. In Figure 4b, the *peak gain* has greatly increased compared to the baseline approach with LS. In Table 1 we list the mean peak gain and Pearson correlation for all seven tracks, along with standard deviations for each value. The first two rows, MS/cos and LS/cos, correspond to the baseline approach—the last three rows summarize the experiments with spectral peak maps. Note that spectral peak maps have approximately ten times the peak gain of MS/LS, whereas the Pearson correlation increases by a factor of four. Figures 3 and 4 illustrate the results in Table 1 at an intuitive level. With MS, the spectral content shared among different types of patterns impedes distinct peaks from emerging. By discarding this irrelevant information, LS better represent the characteristics of each pattern. From the perspective of peak quality, only the self-similarity of the drum pattern continues to pose a challenge.

5. NMFD-BASED EM DECOMPOSITION

By design, our model for EM is very close to the formulation of NMFD; in this section we explore the performance of NMFD and compare it with our fingerprinting methods.

5.1 Related Work

In this section, we briefly review the NMFD method that we employ for decomposing the feature representation V. Weiss and Bello [22] used non-negative matrix factorization (NMF) to identify repeating patterns in music. By adding sparsity constraints and shift-invariant probabilistic latent component analysis (SI-PLCA), they automatically identify the number of patterns and their lengths—applied to beat-synchronous chromagrams in popular music. Masuda et al. [12] propose a query-by-audio system based on NMF to identify the locations where a query musical phrase is present in a musical piece. Among more general techniques for investigating alleged music plagiarism, Dittmar et al. [5] proposed a method for retrieval of sampling. Their approach, based on NMF, was not supplemented with systematic evaluation, but was further investigated in [23]. Previous works [6, 11, 16, 18] successfully applied NMFD—a convolutive version of NMF—for drum transcription and separation. Hockman et al. [7, 8] specifically focused on analyzing breakbeats, i. e., drum-only loops as used in hip hop and drum'n'bass . Detecting sample occurrences throughout a track is a secondary aspect, as they address the more challenging scenario of estimating the loop resequencing [8]. All these previous works have in common that they attempt to retrieve one loop inside a song, whereas we pursue a more holistic approach that allows to deconstruct the whole track into loops.

5.2 NMFD Model

Our objective is to decompose V into component magnitude spectrograms that correspond to the distinct musical elements. Conventional NMF can be used to compute a factorization $\mathrm{V} \approx \mathrm{W} \cdot \mathrm{H}$, where the columns of $\mathrm{W} \in \mathbb{R}_{\geq 0}^{K \times R}$ represent spectral basis functions (also called templates) and the rows of $\mathrm{H} \in \mathbb{R}_{\geq 0}^{R \times M}$ contain time-varying gains (also called activations). The rank $R \in \mathbb{N}$ of the approximation (i. e., number of components) is an important but generally unknown parameter. NMFD extends NMF to the convolutive case by using two-dimensional templates so that each of the R spectral bases can be interpreted as a magnitude spectrogram snippet consisting of $T \ll M$ spectral frames. The convolutive spectrogram approximation $\mathrm{V} \approx \Lambda$ is modeled as

$$\Lambda := \sum_{t=0}^{T-1} \mathrm{W}_t \cdot \overset{t\rightarrow}{\mathrm{H}}, \tag{5}$$

where $\overset{t\rightarrow}{(\cdot)}$ denotes a frame shift operator (see also Eq. 1). As before, each column in $\mathrm{W}_t \in \mathbb{R}_{\geq 0}^{K \times R}$ represents the spectral basis of a particular component, but this time we have T different versions W_t, with $t \in [0 : T-1]$ available. If we take lateral slices along the columns of W_t, we can obtain R prototype magnitude spectrograms $\mathrm{U}^r \in \mathbb{R}_{\geq 0}^{K \times T}$. NMFD typically starts with a suitable initialization (with random values or constant values) of matrices $\mathrm{W}_t^{(0)}$ and $\mathrm{H}^{(0)}$. These matrices are iteratively updated to minimize a suitable distance measure between the convolutive approximation Λ and V. In this work, we use the update rules detailed in [18], which extend the well-known update rules for minimizing the Kullback-Leibler Divergence (KLD) [10] to the convolutive case.

5.3 Evaluation

For our experiments with NMFD we used MS and LS to conduct the procedure in two variants. For the first variant (referred to as R in Table 2), the only *a priori* information used is the number of patterns (or templates) R and their length T. The templates are initialized randomly and

	Gain		*Pearson*	
	μ	σ	μ	σ
NMFD@MS, R	55.20	29.80	0.65	0.25
NMFD@MS, RP	89.62	39.67	0.88	0.11
NMFD@LS, R	52.39	35.95	0.64	0.22
NMFD@LS, RP	79.59	34.99	0.87	0.12

Table 2. Results for NMFD with magnitude spectrograms (MS) and log-frequency spectrograms (LS), initializing the number of templates (R) and also the loop templates (RP). Each column shows the mean and variance for peak gain and Pearson correlation.

fifty iterative updates are used to minimize the KLD. To account for the effects of random initialization, we carry out ten initialization passes per track. The results in Table 2 reflect the mean and standard deviation across all passes. For the second variant (RP), we supply the pattern templates themselves at initialization (i. e., R, T and P are known). We also disallow template updates and only allow activation updates. Since the templates in variant R are initialized randomly, there is no direct relationship between the learned activation curves and the corresponding ground truth curves. We deal with this permutation indeterminacy by comparing all computed activation curves with all ground truth curves and taking the results which maximize the overall score. For all configurations in Table 2, we observe a peak gain at least twice as high as that obtained through diagonal matching; the Pearson correlation increases by a factor of 1.2–1.7, depending on the NMFD configuration taken for comparison. Focusing on the differences among NMFD configurations, RP brings peak gain improvements by a factor slightly greater than 1.5; the Pearson correlation increases by about 1.35. The feature choice (MS or LS) does not play a significant role in result quality. Due to the amount of prior knowledge used to initialize the RP configuration, we consider it as an upper bound for less informed approaches.

6. CONCLUSIONS AND FUTURE WORK

In the preceding sections we developed a better understanding of the feature representations and matching techniques that are commonly used for pattern activation discovery. In this section, we reflect on some of the limitations of our work, further research topics, and computational performance issues.

Clearly, our work only provides a baseline for further work towards more realistic scenarios. As to our model's inherent shortcomings, real-world EM tracks usually contain more than four individual patterns, which are rarely available. Moreover, activations of a given pattern are often (spectrally) different from one another due to the use of effects such as delay and reverb, filter sweeps or resequencing. Thus, we consider this study as a stepping stone towards a fully-developed pipeline for EM structure analysis and decomposition. One potential research direction would be the automatic identification of suitable pattern candidates. A repetition-based analysis technique as described in [14] could be used in conjunction with spectral peak maps to compute self-similarity matrices (SSMs) that saliently encode inclusion relationships. Furthermore, semi-informed variants of NMFD might be helpful in discovering additional patterns that are not explicitly given, where the use of rhythmic structure can serve as prior knowledge to initialize the activations. Although diagonal matching curves can be computed efficiently with a straightforward implementation, we have seen they have certain shortcomings; we wish to investigate the feasibility of using them as rough initial guesses and leaving the refinement up to NMFD. Beyond each method's capabilities, as seen in Tables 1 and 2, there is also the issue of their running time and memory requirements. For the running example, we tested our `MATLAB` implementation on a 3.2 GHz Intel Core i5 CPU with 16 GB RAM, yielding the mean execution times in Table 3. From Tables 3 and 2 we can conclude that NMFD@LS offers the best balance between quality and resource intensity. NMFD@MS takes approximately 14 times longer to compute than NMFD@LS and only produces marginally better results. Indeed, recall that the feature dimensionality $K = 286$ for LS and $K = 2049$ for MS, which explains the large difference in execution times.

Method	*Time (s)*
PLS	0.2
NMFD@LS,(R/RP)	2.5
NMFD@MS,(R/RP)	36.0

Table 3. Computation times for diagonal matching with log-spectral peak maps (PLS), NMFD with magnitude spectrograms (MS), and NMFD with log-frequency spectrograms (LS). The choice of initialization R or RP for NMFD does not impact execution time.

As a final remark, musical structure analysis is an ill-defined problem, primarily because of ambiguity; a segmentation may be based on different principles (homogeneity, repetition, novelty) that can conflict with each other [15]. The main advantage of our method is that we avoid the philosophical issue of how a track's structure is *perceived*, and rather attempt to determine how it was *produced*—a univocal problem. It can then be argued that the listeners' perception is influenced by the cues inherent to EM's compositional style.

7. ACKNOWLEDGMENTS

Patricio López-Serrano is supported in part by CONACYT-DAAD. The authors are supported by the German Research Foundation (DFG MU 2686/6-1, DFG MU 2686/7-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. This work's core ideas were born at HAMR@ISMIR 2015: we thank our teammate Hendrik Schreiber, Colin Raffel, the organizers, and the reviewers for their valuable input.

8. REFERENCES

[1] Ableton. Live. `https://www.ableton.com/en/live/` (web source, retrieved March 2016), 2016.

[2] Jan Balen, Joan Serrà, and Martín Haro. *From Sounds to Music and Emotions: 9th Int. Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers*, chapter Sample Identification in Hip Hop Music, pages 301–312. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[3] Mark J. Butler. *Unlocking the Groove: Rhythm, Meter, and Musical Design in Electronic Dance Music*. Profiles in popular music. Indiana University Press, 2006.

[4] Nick Collins, Margaret Schedel, and Scott Wilson. *Electronic Music*. Cambridge Introductions to Music. Cambridge University Press, Cambridge, United Kingdom, 2013.

[5] Christian Dittmar, Kay Hildebrand, Daniel Gärtner, Manuel Winges, Florian Müller, and Patrick Aichroth. Audio forensics meets music information retrieval - a toolbox for inspection of music plagiarism. In *European Sig. Proc. Conf. (EUSIPCO)*, pages 1249–1253, Bucharest, Romania, August 2012.

[6] Christian Dittmar and Meinard Müller. Towards transient restoration in score-informed audio decomposition. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, pages 145–152, Trondheim, Norway, December 2015.

[7] Jason A. Hockman, Matthew E. P. Davies, and Ichiro Fujinaga. One in the Jungle: Downbeat Detection in Hardcore, Jungle, and Drum and Bass. In *Proc. of the Int. Soc. for Music Inf. Retrieval Conf. (ISMIR)*, pages 169–174, Porto, Portugal, October 2012.

[8] Jason A. Hockman, Matthew E. P. Davies, and Ichiro Fujinaga. Computational strategies for breakbeat classification and resequencing in Hardcore, Jungle and Drum & Bass. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, Trondheim, Norway, December 2015.

[9] Peter Knees, Ángel Faraldo, Perfecto Herrera, Richard Vogl, Sebastian Böck, Florian Hörschläger, and Mickael Le Goff. Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proc. of the Int. Soc. for Music Inf. Retrieval Conf., ISMIR 2015, Málaga, Spain, October 26-30, 2015*, pages 364–370, 2015.

[10] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proc. of the Neural Inf. Processing Systems (NIPS)*, pages 556–562, Denver, CO, USA, 2000.

[11] Henry Lindsay-Smith, Skot McDonald, and Mark Sandler. Drumkit transcription via convolutive NMF. In *Proc. of the Int. Conf. on Digital Audio Effects Conf. (DAFx)*, York, UK, September 2012.

[12] Taro Masuda, Kazuyoshi Yoshii, Masataka Goto, and Shigeo Morishima. Spotting a query phrase from polyphonic music audio signals based on semi-supervised nonnegative matrix factorization. In *Proc. of the Int. Soc. for Music Inf. Retrieval Conf. (ISMIR)*, pages 227–232, Taipei, Taiwan, 2014.

[13] Meinard Müller. *Fundamentals of Music Processing*. Springer Verlag, 2015.

[14] Meinard Müller, Nanzhu Jiang, and Harald Grohganz. SM Toolbox: MATLAB implementations for computing and enhancing similiarty matrices. In *Proc. of the Audio Engineering Soc. AES Conf. on Semantic Audio*, London, UK, 2014.

[15] Jouni Paulus, Meinard Müller, and Anssi P. Klapuri. Audio-based music structure analysis. In *Proc. of the Int. Soc. for Music Inf. Retrieval Conf. (ISMIR)*, pages 625–636, Utrecht, The Netherlands, 2010.

[16] Axel Röbel, Jordi Pons, Marco Liuni, and Mathieu Lagrange. On automatic drum transcription using nonnegative matrix deconvolution and itakura saito divergence. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, pages 414–418, Brisbane, Australia, April 2015.

[17] Bruno Rocha, Niels Bogaards, and Aline Honingh. Segmentation and timbre similarity in electronic dance music. In *Proc. of the Sound and Music Computing Conf. (SMC)*, pages 754–761, Stockholm, Sweden, 2013.

[18] Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Proc. of the Int. Conf. on Independent Component Analysis and Blind Signal Separation ICA*, pages 494–499, Grenada, Spain, 2004.

[19] Rick Snoman. *Dance Music Manual: Tools, Toys, and Techniques*. Taylor & Francis, 2013.

[20] Jan Van Balen. Automatic recognition of samples in musical audio. Master's thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2011.

[21] Avery Wang. An industrial strength audio search algorithm. In *Proc. of the Int. Soc. for Music Inf. Retrieval Conf. (ISMIR)*, pages 7–13, Baltimore, Maryland, USA, 2003.

[22] Ron J. Weiss and Juan Pablo Bello. Unsupervised discovery of temporal structure in music. *IEEE Journal of Selected Topics in Sig. Proc.*, 5:1240–1251, 2011.

[23] Jordan L. Whitney. Automatic recognition of samples in hip-hop music through non-negative matrix factorization. Master's thesis, University of Miami, 2013.