A Repetition-based Triplet Mining Approach for Music Segmentation

Morgan Buisson1,2, Brian McFee3,4, Slim Essid1, Hélène C. Crayencour4

1 LTCI, Télécom Paris, Institut Polytechnique de Paris, France
2 Music and Audio Research Laboratory, New York University, USA
3 Center for Data Science, New York University, USA
4 L2S, CNRS-Univ Paris-Sud-CentraleSupélec, France

Contributions

The notion of repetition is tightly linked to musical structures and has not been explicitly considered by previous self-supervised methods for audio representation learning and music segmentation. In this work we propose:

• An unsupervised triplet mining method to learn audio representations for music segmentation.
• We leverage repeating sequences inside the input track to select relevant sets of frames to learn a deep neural network with a triplet loss.
• Our approach returns more informative triplets, which enhances the learned representations.
• The output embeddings significantly improve both boundary detection and section grouping results against comparable previous work.
• We provide further insight on the relationship between the nature of the repetitions leveraged and the music genre employed for testing.

Method Overview

For each track in the training set (non-annotated):

1. Calculate MFCC and Chroma features, convert to time-lag features
2. Calculate their respective self-similarity matrices \( S_M \) and \( S_C \).
3. Linear combination matrix \( S = \gamma S_M + (1-\gamma)S_C, \gamma \in [0,1] \).
4. Filtering operation and diagonal enhancement of \( S \).
5. Dilation operation on \( S \) to obtain the positive sampling matrix \( S_P \).
6. Exponential decay on \( 1 - S_P \) to obtain the negative sampling matrix \( S_N \).
7. For any frame randomly chosen, select positive example by querying the matrix \( S_P \) and the negative with \( S_N \).

Time-lag self-similarity matrices

![Figure 1. Time-lag self-similarity matrices \( S_M \) and \( S_C \), respectively.](image)

\[
M(i,j) = \begin{cases} 
\exp \left( -\frac{d(x_i, y_j)^2}{\lambda} \right), & x_i, y_j \in NN_M(x_i), \\
0, & x_i \notin NN_M(x_i) 
\end{cases}
\]

Combination of timbral and harmonic repetitions

The matrix \( S \) is then obtained by linear combination, such that:

\[
S = \gamma S_M + (1-\gamma)S_C,
\]

where \( \gamma \in [0,1] \) weights the contributions of each feature type.

![Figure 2. Repetition matrix \( S \) (left) and reference self-similarity matrix (right).](image)

We first set \( \gamma = 0.5 \); equal weight to timbral and harmonic features is given.

Sampling matrices

Positive sampling matrix: A dilation operation is applied to the matrix \( S \) to enlarge these detected regions of repetition. A two-dimensional Gaussian kernel \( G \) of size \( K \) is convolved with \( S \):

\[
S_P = S \ast G,
\]

The size of the kernel \( K \) was set to \( K = 8 \) (beats), providing a good balance between the amount of dilation and its alignment with segment boundaries, as it blurs repetitions over 2 bars when songs follow a 4/4 time signature.

Negative sampling matrix: The negative sampling matrix \( S_N \) is obtained by applying an exponential decay to \( 1 - S_P \) such that:

\[
S_N(i,j) = (1 - S_P(i,j))e^{-\lambda (\|x_i - x_j\|)},
\]

where \( \lambda > 0 \) is a parameter that defines the strength of the smoothing. Components near the main diagonal of \( S_N \) receive greater values than those close the opposite edges, thus favoring frames located within consecutive segments of that of the anchor.

Segmentation evaluation

• Training set: 20,000 non annotated audio tracks, splits of 16%, 50% and 100%.
• Test datasets: SALAMI [1] and JSD [2].
• Downstream segmentation and section grouping performed with spectral clustering [3].
• Evaluation metrics: HR.5F, HR.3F (Hit-Rate F-measures with .5 and 3 second tolerance windows), F-measure of frame pairwise clustering (PPC) and normalized conditional entropy score (NCE).
• Baselines: spectral clustering [3] applied on positive sampling matrix (LSD) and temporal sampling [4].

<table>
<thead>
<tr>
<th>Method Split</th>
<th>LSD</th>
<th>HR.5F</th>
<th>HR.3F</th>
<th>PPC</th>
<th>NCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp. (10%)</td>
<td>303</td>
<td>535</td>
<td>467</td>
<td>677</td>
<td>678</td>
</tr>
<tr>
<td>Ours (10%)</td>
<td>291</td>
<td>676</td>
<td>773</td>
<td>678</td>
<td>690</td>
</tr>
<tr>
<td>Temp. (50%)</td>
<td>284</td>
<td>671</td>
<td>773</td>
<td>678</td>
<td>690</td>
</tr>
<tr>
<td>Ours (50%)</td>
<td>296</td>
<td>682</td>
<td>778</td>
<td>690</td>
<td>694</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method Split</th>
<th>LSD</th>
<th>HR.5F</th>
<th>HR.3F</th>
<th>PPC</th>
<th>NCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp. (10%)</td>
<td>243</td>
<td>670</td>
<td>773</td>
<td>678</td>
<td>690</td>
</tr>
<tr>
<td>Ours (10%)</td>
<td>297</td>
<td>683</td>
<td>781</td>
<td>694</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Flat segmentation results on SALAMI (upper annotations). Results in bold denote statistically significant improvement over temporal sampling on same split (denoted as Temp.).

<table>
<thead>
<tr>
<th>Method Split</th>
<th>LSD</th>
<th>HR.5F</th>
<th>HR.3F</th>
<th>PPC</th>
<th>NCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp. (10%)</td>
<td>228</td>
<td>686</td>
<td>788</td>
<td>707</td>
<td>712</td>
</tr>
<tr>
<td>Ours (10%)</td>
<td>223</td>
<td>685</td>
<td>789</td>
<td>709</td>
<td>717</td>
</tr>
<tr>
<td>Temp. (50%)</td>
<td>231</td>
<td>695</td>
<td>793</td>
<td>712</td>
<td>712</td>
</tr>
<tr>
<td>Ours (50%)</td>
<td>234</td>
<td>696</td>
<td>793</td>
<td>712</td>
<td>712</td>
</tr>
</tbody>
</table>

Table 2: Flat segmentation results on JSD (chorus annotation level) with emphasis on timbral features \( \gamma = 0.5 \). Results in bold denote statistically significant improvement over temporal sampling (denoted as Temp.) on same split.

Conclusion

• Boundary detection and structural grouping improved in a significant manner on all splits.
• Better triplets generated, improves the training signal and convergence.
• Influence of balance parameter \( \gamma \):
  - Emphasizing timbral content: “non-repeating” structure types (Jazz).
  - More weight on harmonic content: better for repetition-based structures (Pop, Rock).

References