A Repetition-based Triplet Mining Approach for Music Segmentation

Morgan Buisson[1,4], Brian McFee[2,3], Slim Essid[1], Hélène C. Crayencour[4]

[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 train 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 $\mathbf{S}_M$ and $\mathbf{S}_C$
3. Linear combination matrix $\mathbf{S} = \gamma\mathbf{S}_M + (1-\gamma)\mathbf{S}_C, \gamma \in [0,1]$
4. Filtering operation and diagonal enhancement of $\mathbf{S}$
5. Dilation operation on $\mathbf{S}$ to obtain the positive sampling matrix $\mathbf{S}_P$
6. Exponential decay on $1-\mathbf{S}_P$ to obtain the negative sampling matrix $\mathbf{S}_N$
7. For any frame randomly chosen, select positive example by querying the matrix $\mathbf{S}_P$ and the negative with $\mathbf{S}_N$

Time-lag self-similarity matrices

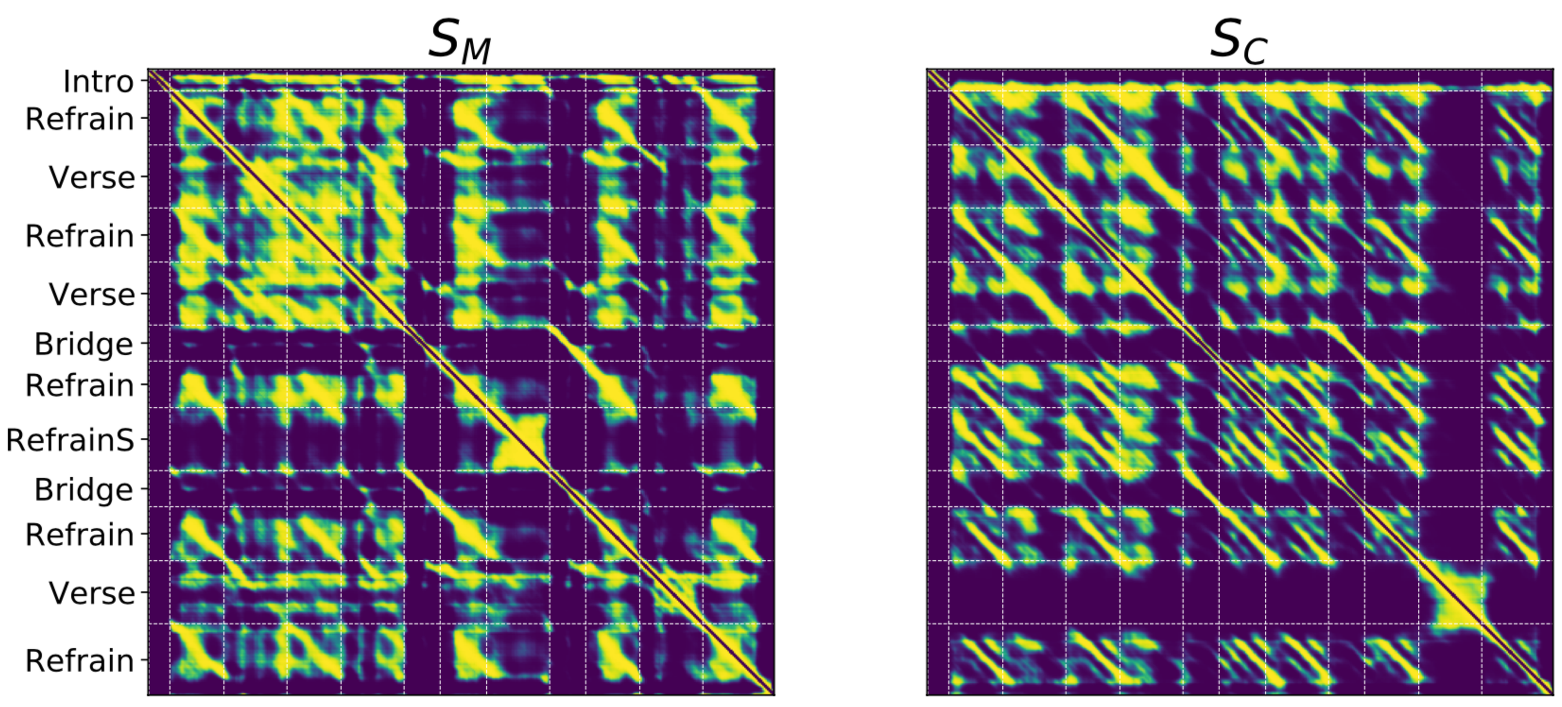

Figure 1. Time-lag self-similarity matrices $\mathbf{S}_M$ and $\mathbf{S}_C$ respectively.

$$M(i,j) = \begin{cases} \exp\left(-\frac{d(\tilde{\mathbf{X}}_i,\tilde{\mathbf{X}}_j)}{b}\right), & \tilde{\mathbf{X}}_j \in \mathrm{NN_k}(\tilde{\mathbf{X}}_i) \\ 0, & \tilde{\mathbf{X}}_j \notin \mathrm{NN_k}(\tilde{\mathbf{X}}_i) \end{cases} \tag{1}$$

where $\tilde{\mathbf{X}}_i$ are time-lag features, $d(x,y)$ is the euclidean distance, b the bandwidth parameter, $\mathrm{NN_k}(x)$ denotes the k-nearest neighbors of x and $i,j = 1,\ldots,N$.

Combination of timbral and harmonic repetitions

The matrix $\mathbf{S}$ is then obtained by linear combination, such that:

$$\mathbf{S} = \gamma\mathbf{S}_M + (1-\gamma)\mathbf{S}_C, \tag{2}$$

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

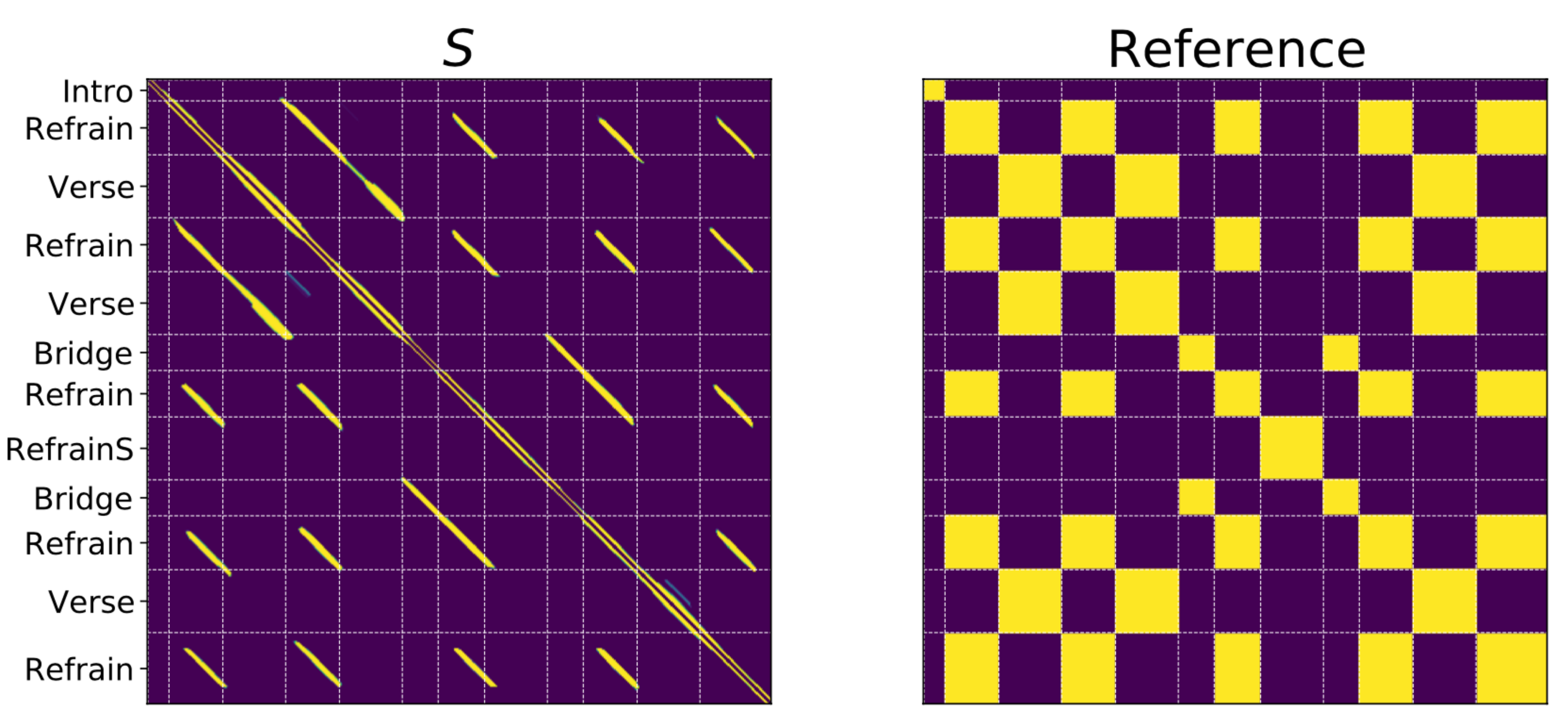

Figure 2. Repetition matrix $\mathbf{S}$ (left) and reference self-similarity matrix (right).

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 $\mathbf{S}$ to enlarge these detected regions of repetition. A two-dimensional Gaussian kernel G of size K is convolved with $\mathbf{S}$:

$$\mathbf{S}_P = \mathbf{S} * G, \tag{3}$$

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 $\mathbf{S}_N$ is obtained by applying an exponential decay to $1-\mathbf{S}_P$ such that:

$$\mathbf{S}_N(i,j) = (1-\mathbf{S}_P(i,j))e^{-\lambda \max\left(\frac{|i-j|}{N}, \mathbf{S}_P(i,j)\right)}, \tag{4}$$

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

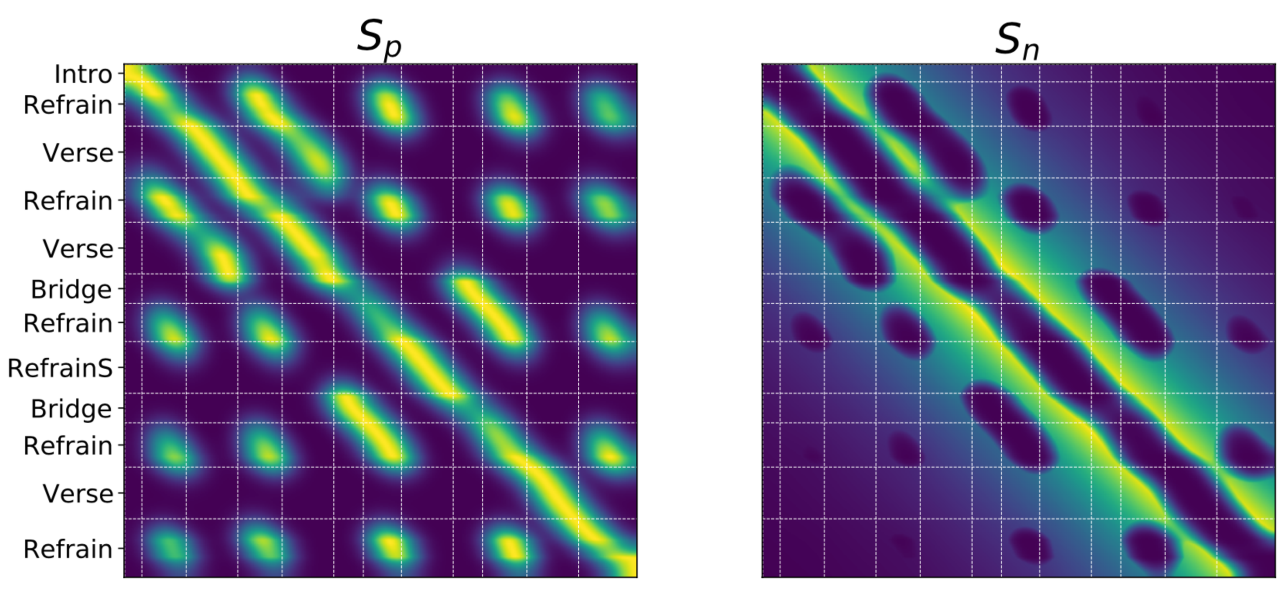

Figure 3. Positive (left) and negative sampling matrices (right) $\mathbf{S}_P$ and $\mathbf{S}_N$.

Segmentation evaluation

- **Training set**: $20,000$ non annotated audio tracks, splits of 10%, 50% and 100%.
- **Test datasets**: SALAMI [1] and JSD [2].
- Downstream segmentation and section grouping performed with spectral clustering [3].
- **Evaluation metrics**: HR.5F, HR3F (Hit-Rate F-measures with .5 and 3 second tolerance windows), F-measure of frame pairwise clustering (PFC) and normalized conditional entropy score (NCE).
- **Baselines**: spectral clustering [3] applied on positive sampling matrix (LSD) and temporal sampling [4].

Method (Split)	HR.5F	HR3F	PFC	NCE
LSD	.195	.486	.707	.682
Temp. (10%) [4]	.280	.665	.770	.677
Ours (10%)	**.291**	**.676**	**.777**	**.691**
Temp. (50%) [4]	.288	.671	.773	.678
Ours (50%)	**.296**	**.682**	**.778**	**.690**
Temp. (100%) [4]	.284	.670	.773	.678
Ours (100%)	**.297**	**.683**	**.781**	**.694**

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.*).

Method (Split)	HR.5F	HR3F	PFC	NCE
Temp. (10%) [4]	.221	.568	.739	.745
Ours (10%, $\gamma = 0.9$)	.223	**.585**	.743	.750
Temp. (50%) [4]	.243	.586	.763	.766
Ours (50%, $\gamma = 0.9$)	.234	**.607**	.769	.772

Table 2. Flat segmentation results on JSD (*chorus* annotation level) with emphasis on timbral features ($\gamma = 0.9$). 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 γ:
  - Emphasizing **timbral** content: "**non-repeating**" structure types (Jazz).
  - More weight on **harmonic** content: better for **repetition-based** structures (Pop, Rock).

References

[1] Jordan Bennett Louis Smith et al. "Design and creation of a large-scale database of structural annotations.". In: *ISMIR*. 2011.

[2] Stefan Balke et al. "JSD: A Dataset for Structure Analysis in Jazz Music". In: *Transactions of the International Society for Music Information Retrieval* 5.1 (2022).

[3] Brian McFee and Dan Ellis. "Analyzing Song Structure with Spectral Clustering.". In: *ISMIR*. 2014.

[4] Matthew C McCallum. "Unsupervised learning of deep features for music segmentation". In: *ICASSP*. 2019.

morgan.buisson@telecom-paris.fr