Convolutive Block-Matching Segmentation Algorithm with Application to Music Structure Analysis

A. Marmoret[1,2], J.E. Cohen[3], F. Bimbot[1]

[1] Univ Rennes, Inria, CNRS, IRISA, France, [2] IMT Atlantique, Lab-STICC, Brest, France, [3] CNRS, CREATIS, Villeurbanne France

axel.marmoret@imt-atlantique.fr

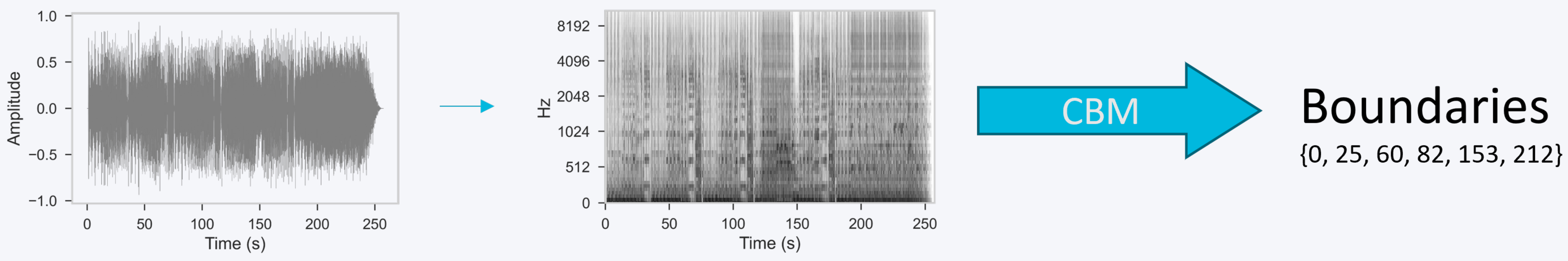
Poster summary

This poster presents an algorithm aiming at segmenting autosimilarity matrices, called Convolutive Block-Matching (CBM) algorithm.

The CBM algorithm aims at framing blocks of high self-similarity in an autosimilarity matrix, *i.e.* homogeneous regions.

The CBM is introduced for the task of Music Structure Analysis (MSA), by segmenting songs sampled at the barscale.

The proposed algorithm achieves a level of performance competitive to that of supervised State-of-the-Art methods on 3 among 4 metrics while being unsupervised.



Barwise processing

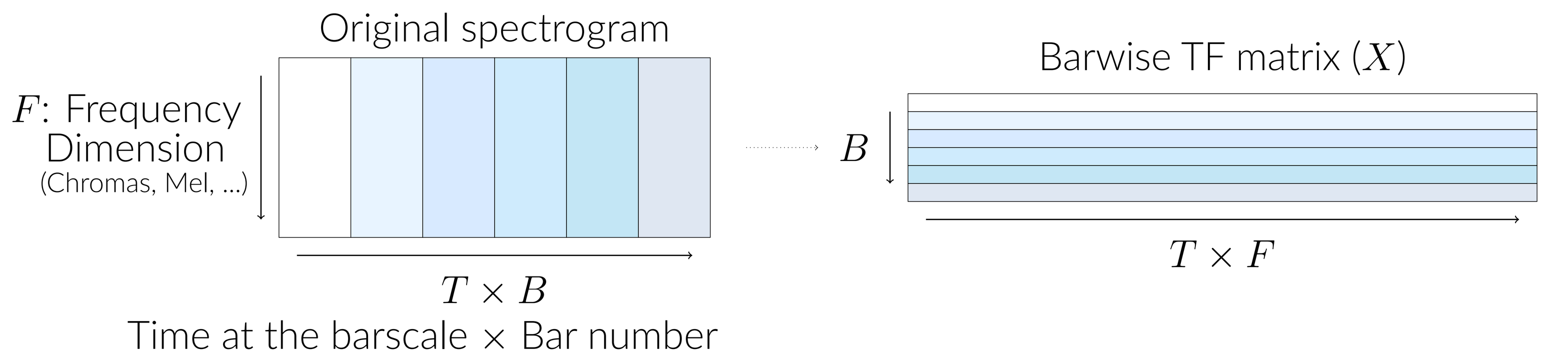


Figure 1. The spectrogram is cut at each downbeat, and the information contained in each bar is vectorized. This results in a **Barwise TF matrix**, of size $B \times TF$.

Autosimilarity matrix

An autosimilarity matrix $A(X) \in \mathbb{R}^{B \times B}$ contains the similarity between all pair of bars:

$$A(X)_{ij} = s(X_i, X_j) \tag{1}$$

Figure 2. The spectrogram is cut on each downbeat, and the information contained in each bar is vectorized. This results in a **Barwise TF matrix**, of size $B \times TF$.

Three similarity functions are studied here:

$$s(X_i, X_j) = \begin{cases} \text{Cosine similarity} & : \frac{\langle X_i, X_j \rangle}{\|X_i\|_2 \|X_j\|_2} \\ \text{Covariance similarity} & : \frac{\langle X_i - \bar{x}, X_j - \bar{x} \rangle}{\|X_i - \bar{x}\|_2 \|X_j - \bar{x}\|_2} \\ \text{RBF similarity} & : \exp\left(-\gamma \left\| \frac{X_i}{\|X_i\|_2} - \frac{X_j}{\|X_j\|_2} \right\|_2^2\right) \end{cases} \tag{2}$$

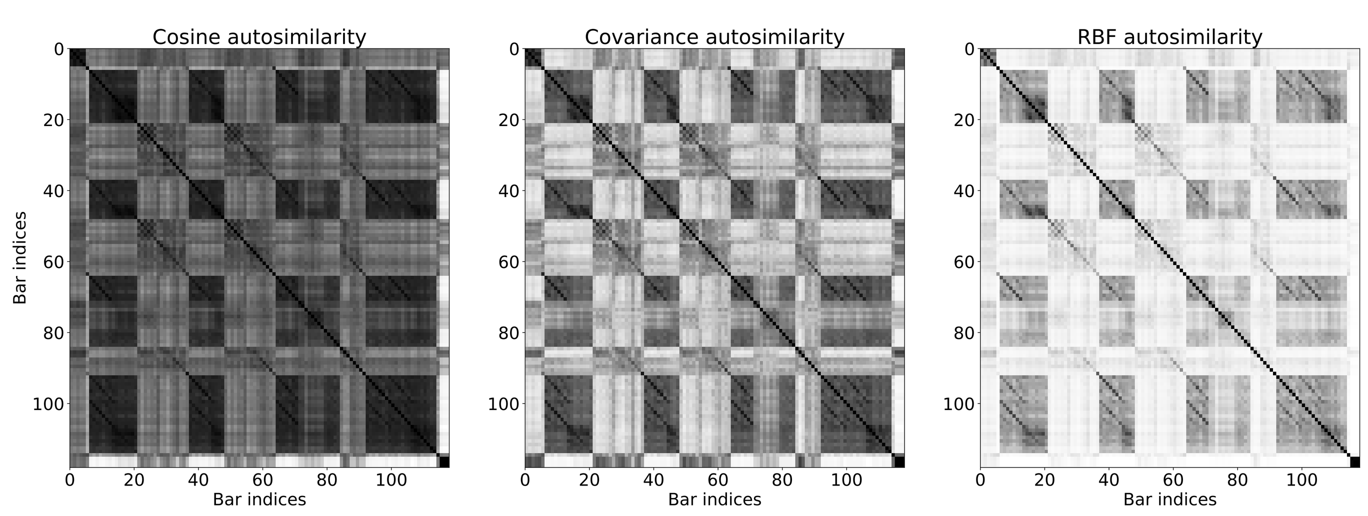


Figure 3. Cosine, Covariance and RBF autosimilarities on the song *POP01* from RWC Pop.

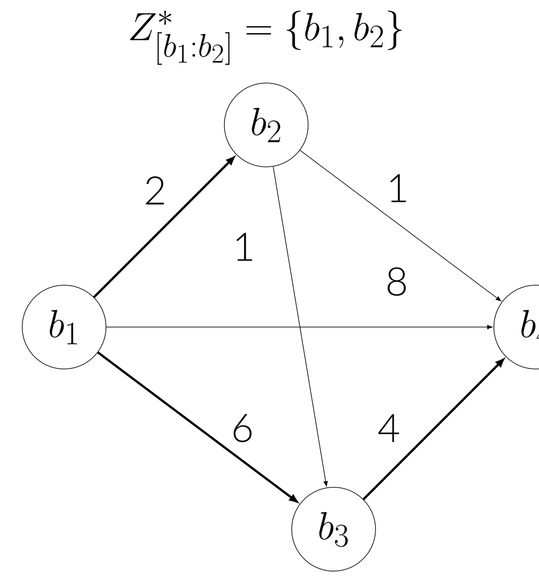
Algorithm principles

Notations:

- $Z^*_{[b_i:b_j]}$: optimal segmentation (set of boundaries) between bars b_i and b_j.
- $u()$: score function (for a segment or a set of segments).

Framed as a Longest-path in a graph (directed and acyclic)

$$Z^*_{[b_1:b_2]} = \{b_1, b_2\}$$



$$Z^*_{[b_1:b_4]} = Z^*_{[b_1:b_3]} \cup \{b_4\} = \{b_1, b_3, b_4\}$$

$$Z^*_{[b_1:b_3]} = \{b_1, b_3\}$$

Solution: $u\left(Z^*_{[b_1:b_3]}\right) = \max\left(u\left(Z^*_{[b_1:b_2]}\right) + u\left([\![b_2, b_3]\!]\right), u\left([\![b_1, b_3]\!]\right)\right) = 6$

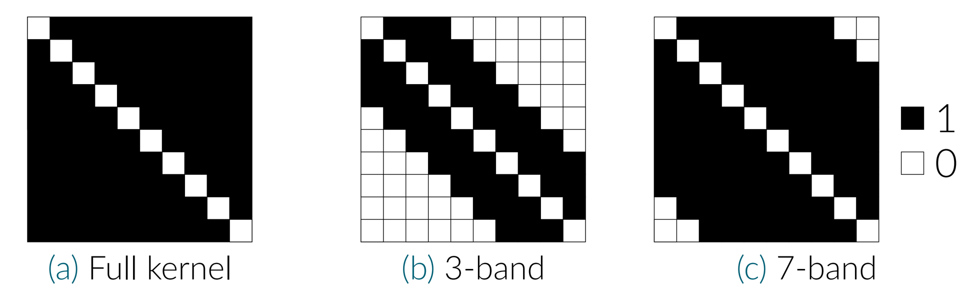
Formally, this is written as an optimization problem, depending on the function u:

$$Z^* = \underset{Z \in \Theta}{\arg\max} \sum_{i=1}^{E-1} u(S_i). \tag{3}$$

Score function

$$u(S_i) = \frac{1}{\nu |S_i|} \sum_{k=1}^{|S_i|} \sum_{l=1}^{|S_i|} A_{S_i}(X)_{kl} K_{kl} - \lambda p(|S_i|). \tag{4}$$

Convolution kernels (block-weighting) — Penalty function



(a) Full kernel (b) 3-band (c) 7-band

Figure 4. Different kernels, of size 10

$$p(|S_i|) = \begin{cases} 0 & \text{if } |S_i| = 8 \\ \frac{1}{4} & \text{else if } |S_i| \equiv 0 \pmod 4 \\ \frac{1}{2} & \text{else if } |S_i| \equiv 0 \pmod 2 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

Quantitative results

Results according to parameters of the CBM:

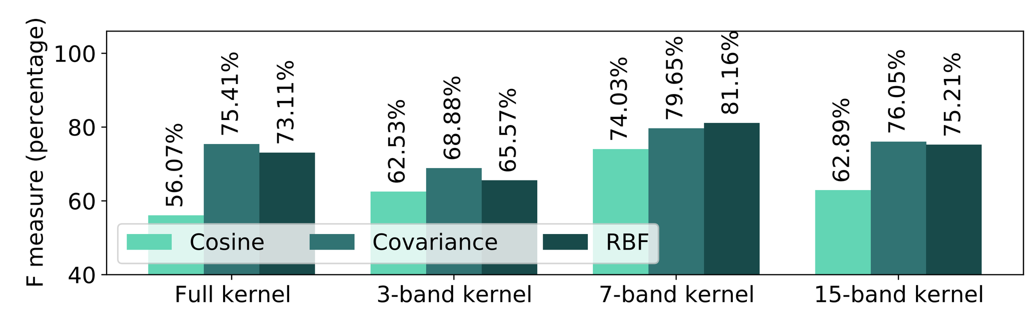


Figure 5. Results according to the similarity function and convolution kernel. F_{3s} on the RWC Pop dataset.

Best results, compared with State-of-the-Art algorithms (SOTA) [1, 2, 3, 4, 5, 6, 7]

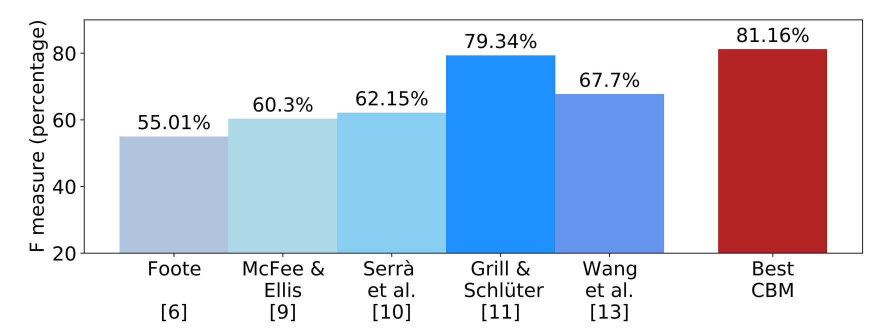


Figure 6. F_{3s} on the RWC Pop dataset.

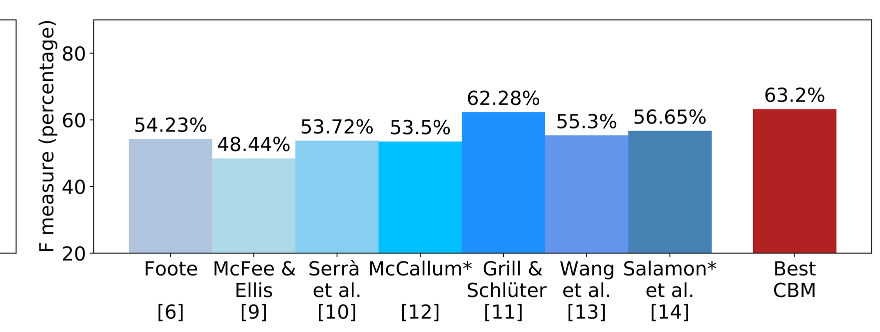


Figure 7. F_{3s} on the SALAMI dataset.

More detailed results on the paper!

Open-source toolbox

https://gitlab.imt-atlantique.fr/a23marmo/autosimilarity_segmentation/-/tree/WASPAA23

Take home messages

1. A new segmentation algorithm!
   - High performances, without supervision (still necessitates downbeat estimation)
   - **Low-complexity** and **easily customizable**,
2. May be used with **any representation-learning algorithm** (*e.g.* your favourite neural network).
3. **Barwise sampling** participates in boosting the performance of music structure estimation (more experiments in the future detailed version).

Perspectives (contact me! :))

1. Studying (or learning) different types of kernels,
2. Improving the penalty function (empirical),
3. Replace the simple similarity functions with more complex ones (*e.g.* learned by means of a neural network).

References

[1] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *IEEE International Conference on Multimedia and Expo. Proceedings Latest Advances in the Fast Changing World of Multimedia*, pp. 452–455, IEEE, 2000.

[2] B. McFee and D. Ellis, "Analyzing song structure with spectral clustering," in *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 405–410, 2014.

[3] J. Serra, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.

[4] M. C. McCallum, "Unsupervised learning of deep features for music segmentation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 346–350, IEEE, 2019.

[5] T. Grill and J. Schlüter, "Music boundary detection using neural networks on combined features and two-level annotations," in *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 531–537, 2015.

[6] J.-C. Wang, J. B. Smith, W.-T. Lu, and X. Song, "Supervised metric learning for music structure feature," in *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 730–737, 2021.

[7] J. Salamon, O. Nieto, and N. J. Bryan, "Deep embeddings and section fusion improve music segmentation," in *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 594–601, 2021.