LEARNING MULTI-LEVEL REPRESENTATIONS FOR HIERARCHICAL MUSIC STRUCTURE ANALYSIS

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 of Data Science, New York University, USA
[4] L2S, CNRS-Univ.Paris-Sud-CentraleSupélec, France

Contributions

- **Multi-level triplet mining** method requiring no structural annotations.
- **Disentangled representation learning** to extract deep features at different levels of granularity.
- Method competitive against related work for both **single-level** and **multi-level** music segmentation.
- Good **generalization** across annotators, segmentation levels and various datasets for multi-level segmentation.

Learning audio representations from triplets

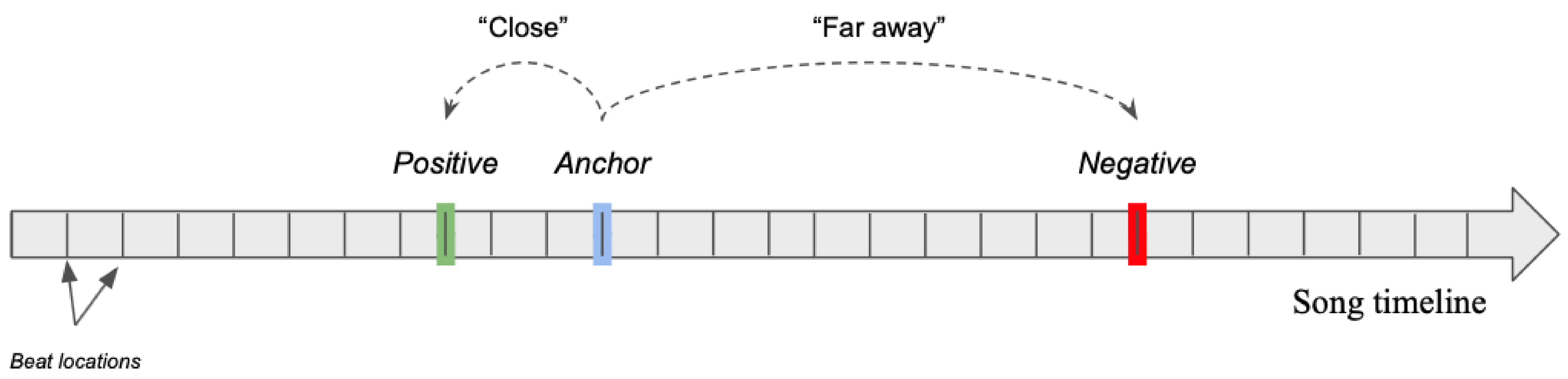
For a triplet of frames $\mathcal{T} = (x_a, x_p, x_n)$, the triplet loss is expressed as:

$$\mathcal{L}(\mathcal{T}) = [d(f(x_a), f(x_p)) - d(f(x_a), f(x_n)) + \alpha]_+,$$

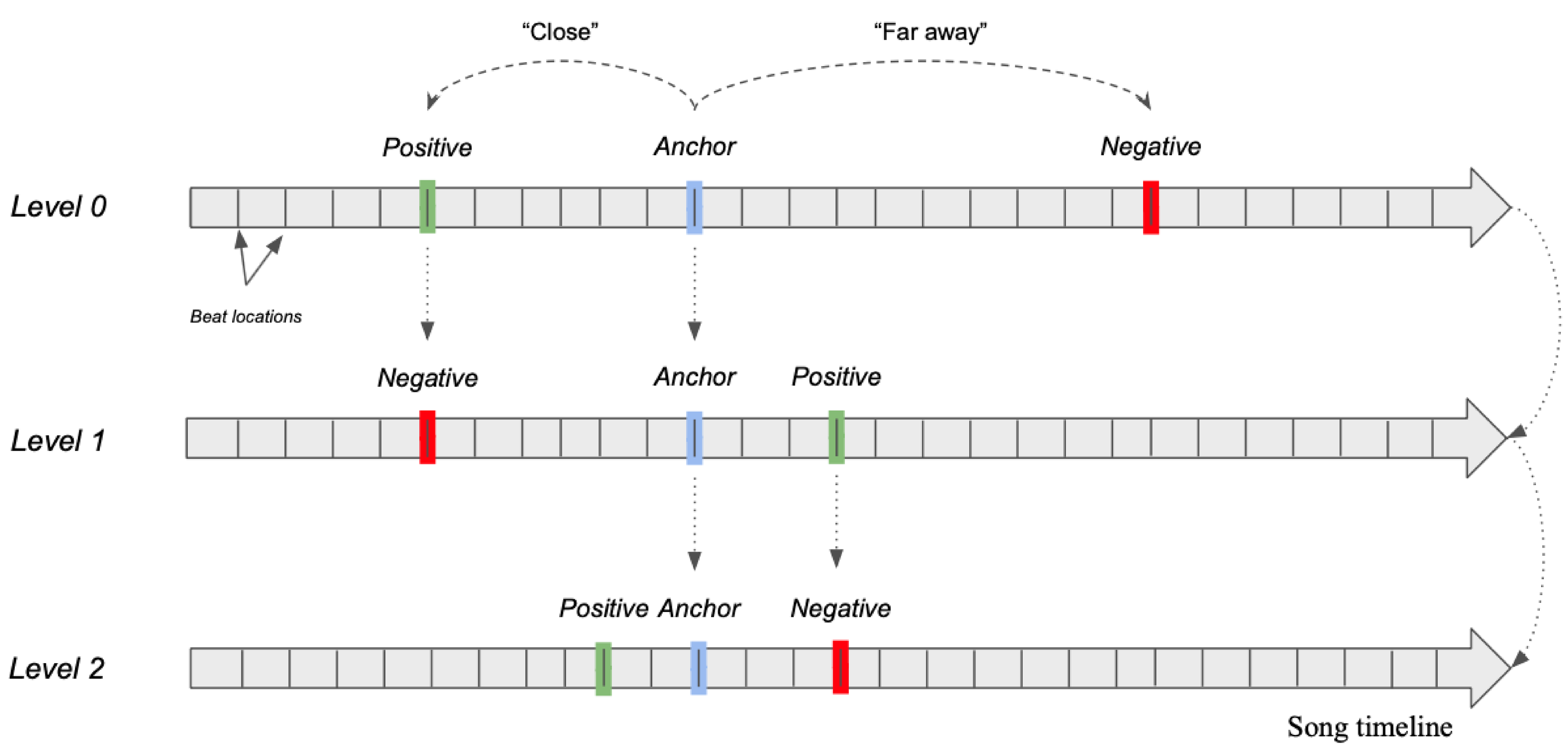
where $d(x, y)$ is a distance metric, $[.]_+$ the Hinge loss, α the margin parameter and $f(x)$ the projection of x into the embedding space by a deep neural network.

Sampling triplets without annotations

A first triplet mining method approach has recently been proposed [1]:



The method introduced here can be viewed as its **multi-level extension**:



For a randomly sampled anchor index at level $\ell = 0$, a complete triplet (x_a, x_p, x_n) is built only using time proximity as a proxy. For each level $\ell \in \{1;\dots;L-1\}$, the positive example is sampled closer and closer to the same anchor, whereas the negative is obtained by selecting the positive example from level $\ell - 1$.

Disentangling hierarchy levels

Triplets sampled at different hierarchy levels used to train sub-regions of the output embeddings [2]. A set of L masking functions $m_\ell \in \{0, 1\}^n$ that are applied to the embedding space of size n is defined. For a given triplet (x_a, x_p, x_n) at level ℓ, the training objective becomes:

$$\mathcal{L}(x_a, x_p, x_n) = [D_\ell(x_a, x_p) - D_\ell(x_a, x_n) + \alpha]_+,$$

where:

$$D_\ell(x_i, x_j) = \| m_\ell \circ [f(x_i) - f(x_j)] \|_2^2$$

Details:

- L is fixed beforehand.
- Mining step done offline.
- Masks initialized with fixed and equal length.
- $\forall \ell \in \{0;\dots;L-2\}, \alpha_\ell > \alpha_{\ell+1}$.

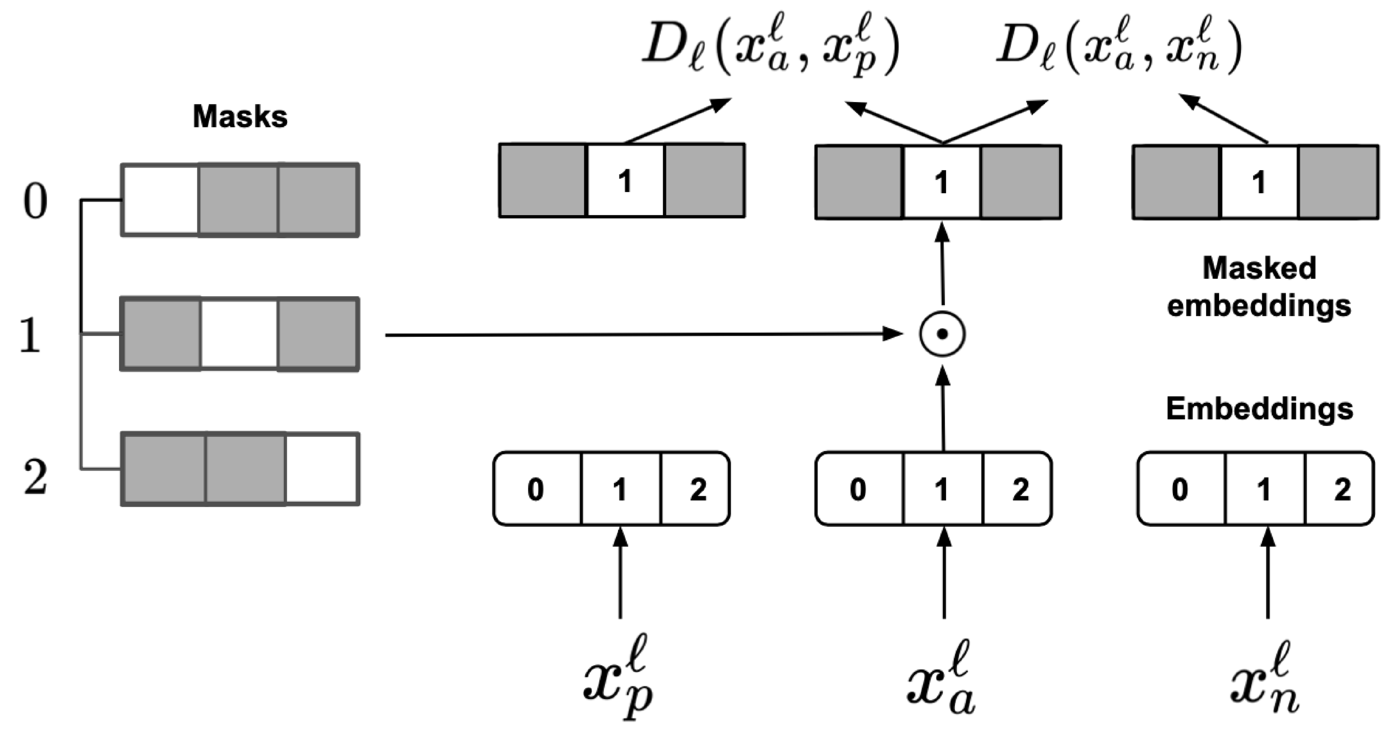


Figure 1: Training pipeline for a triplet of frames at level ℓ.

Experiments

- **Training data**: $23,725$ non annotated tracks spanning various musical genres.
- **Testing data**: BeatlesTUT, RWC-Pop, RWC-Jazz, JAAH and a doubly annotated subset of SALAMI. Hierarchy expansion has been applied to all datasets [3].
- **Metrics**: We report the F-measure of the trimmed boundary detection hit-rate with a 3-second tolerance window (F_3), the F-measure of frame pairwise clustering (F_{pairwise}) and the L-measure [4] for multi-level segmentation.

Results

Level	*lower*					*upper*						*combined*	
Method	LSD	FE_0	HE_0	FE_1	HE_1	SNF	LSD	FE_0	HE_0	FE_1	HE_1	HE_0	HE_1
F_3	0.525	**0.624**	0.611	**0.611**	0.600	0.456	0.579	0.568	**0.597**	0.559	**0.595**	0.665	0.662
F_{pairwise}	0.561	0.561	**0.580**	0.563	**0.581**	0.567	0.652	0.694	**0.714**	0.697	**0.718**	0.730	0.731

Table 1: Boundary detection and section grouping results on SALAMI. FE: flat embeddings [1]. HE: hierarchical embeddings (ours).

- Boundary detection improved at *upper* level for both annotators.
- Frame-wise assignment performance increased for both annotators at both level of granularity.

Method	Inter-annot	LSD	SNF	DEF	FE_0	HE_0	FE_1	HE_1
L-Precision	0.664	0.419	0.431	0.435	0.412	**0.413**	0.413	**0.418**
L-Recall	0.664	0.636	0.668	0.673	0.677	**0.680**	0.663	**0.671**
L-Measure	0.654	0.498	0.517	0.520	0.505	**0.507**	0.503	**0.509**

Table 2: Multi-level segmentation results on SALAMI. Inter-annot denotes the inter-annotator agreement.

- Evaluation focused on L-Recall values: representing how much of the reference hierarchy is retrieved by the predicted one.
- Reference hierarchies better retrieved by hierarchical representations than baselines employing multiple features or operating at only one level of granularity.

Dataset	F_3	F_{pairwise}	L-P	L-R	L-M
BeatlesTUT	71.77	72.25	49.32	75.25	59.37
RWC-Pop	68.07	65.35	47.02	77.06	58.30
RWC-Jazz	55.05	58.51	32.89	81.80	45.76
JAAH	55.57	76.72	46.49	81.18	58.55

Table 3: Boundary detection, section grouping and multi-level segmentation results on additional datasets (in percentage) with the whole embedding matrix. L-P: L-precision, L-R: L-recall, L-M: L-measure.

- Mixed performance in terms of boundary detection and section grouping: due to varying annotation levels and music genres.
- The L-recall values obtained across datasets remain within the same range: most of the reference structure hierarchies are captured.

References

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

[2] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. "Conditional similarity networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[3] Brian McFee and Katherine M Kinnaird. "Improving structure evaluation through automatic hierarchy expansion". In: *ISMIR*. 2019.

[4] Brian McFee et al. "Evaluating hierarchical structure in music annotations". In: *Frontiers in psychology* (2017).