Extended Paper. Updated February 27, 2014

A long-range self-similarity approach to segmenting DJ mixed music streams

Tim Scarfe TIM@DEVELOPER-X.COM
Wouter M. Koolen WOUTER@CS.RHUL.AC.UK
Yuri Kalnishkan YURA@CS.RHUL.AC.UK
Computer Learning Research Centre and Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, United Kingdom

Abstract

We describe an unsupervised, deterministic algorithm for segmenting DJ-mixed Electronic Dance Music (EDM) streams (for example; pod-casts, radio shows, live events) into their respective tracks. We attempt to reconstruct boundaries as close as possible to what a human domain expert would create in respect of the same task. The goal of DJ-mixing is to render track boundaries effectively invisible from the standpoint of human perception which makes the problem difficult.

We use dynamic programming to optimally segment a cost matrix derived from a similarity matrix. The similarity matrix is based on the cosines of a time series of kernel-transformed Fourier based features designed with this domain in mind. Our method is applied to EDM streams. Its formulation incorporates long-term self similarity as a first class concept combined with dynamic programming and it is qualitatively assessed on a large corpus of long streams that have been hand labelled by a domain expert.

Keywords: music segmentation DJ mix dynamic programming

1. Cost Matrix

We now have a dissimilarity matrix $S(i,j)$ as described in Section **??**.

Let w and W denote the minimum and maximum track length in seconds, these will be parameters.

Intuitively, features within the same track are reasonably similar on the whole, while pairs of tiles that do not belong to the same track are significantly more dissimilar. We define $C(f,t)$, the cost of a candidate track from tile f through tile t, to be the sum of the dissimilarities between all pairs of tiles inside it, normalized on track length:

$$C(f,t) \;=\; \frac{\sum_{i=f}^{t}\sum_{j=f}^{t} S(i,j)}{\sqrt{t-f+1}}$$

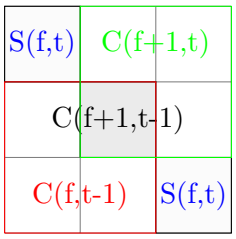
As a first step, we pre-compute C for each $1 \le f \le t \le T$. Direct calculation using the definition takes $O(TW^3)$ time. However, we can compute the full cost matrix in $O(WT)$ time using the following recursion for the unnormalized quantity $\tilde{C}(f,t) = C(f,t)(t-f)$ (for $f+1 \le t-1$)

$$\tilde{C}(f,t) \;=\; \tilde{C}(f+1,t) + \tilde{C}(f,t-1) - \tilde{C}(f+1,t-1) + S(f,t) + S(t,f).$$

© T. Scarfe, W.M. Koolen & Y. Kalnishkan. Artificial Intelligence Applications and Innovations.

Note that the normalization step can be done independently of the DP procedure. We discovered experimentally that normalizing using the square root of the track length was advantageous. Doing so slightly discourages tracks of a larger length.

Here is a visualisation of the update routine in the domain of S.



2. Computing Best Segmentation

We obtain the cost of a full segmentation by summing the costs of its tracks. The goal is now to efficiently compute the segmentation of least cost.

A sequence $\boldsymbol{t} = (t_1, \ldots, t_{m+1})$ is called an m/T-segmentation if

$$1 = t_1 < \ldots < t_m < t_{m+1} = T + 1.$$

m is the number of tracks we are trying to find and is a parameter of the algorithm. We use the interpretation that track $i \in \{1, \ldots, m\}$ comprises times $\{t_i, \ldots, t_{i+1} - 1\}$. Let $\mathbb{S}_m^T$ be the set of all m/T-segmentations. Note that there is a very large number of possible segmentations

$$|\mathbb{S}_m^T| = \binom{T-1}{m-1} = \frac{(T-1)!}{(m-1)!(T-m)!} = \frac{(T-1)(T-2)\cdots(T-m+1)}{(m-1)!} \geq \left(\frac{T}{m}\right)^{m-1}.$$

For large values of T, considering all possible segmentations using brute force is infeasible. For example, a two hour long show with 25 tracks would have more than $\left(\frac{60^2 \times 2}{25}\right)^{24} \approx 1.06 \times 10^{59}$ possible segmentations!

We can reduce this number slightly by imposing upper and lower bounds on the song length. Recall that W is the upper bound (in seconds) of the song length, w the lower bound (in seconds) and m the number of tracks. With the track length restriction in place, the number of possible segmentations is still massive. A number now on the order of 10^{56} for a two hour show with 25 tracks, $w = 190$ and $W = 60 \times 15$.

Let $N(T, W, w, m)$ be the number of segmentations with time T (in tiles),

We can write the recursive relation

$$N(T, W, w, m) = \sum N(t_m - 1, W, w, m - 1)$$

, where the sum is taken over t_m such that

$$t_m \leq T - w + 1 \qquad t_m \geq T - W + 1$$
$$t_m \geq (m-1)w + 1 \qquad t_m \leq (m-1)W + 1$$

The first two inequalities mean that the length of the last track is within an acceptable boundary between w and W. The last two inequalities mean that the lengths of the first $m-1$ tracks are within the same boundaries.

We calculated the value of $N(7000, 60 \times 15, 190, 25)$ and got 5.20×10^{56} which is still infeasible to compute with brute force.

Our solution to this problem is to find a dynamic programming recursion.

The loss of an m/T-segmentation $\boldsymbol{t}$ is

$$\ell(\boldsymbol{t}) \;=\; \sum_{i=1}^{m} C(t_i, t_{i+1}-1)$$

We want to compute

$$\mathcal{V}_m^T \;=\; \min_{\boldsymbol{t}\in\mathbb{S}_m^T} \ell(\boldsymbol{t})$$

To this end, we write the recurrence

$$\mathcal{V}_1^t \;=\; C(1,t)$$

and for $i \geq 2$

$$\mathcal{V}_i^t \;=\; \min_{\boldsymbol{t}\in\mathbb{S}_i^t} \ell(\boldsymbol{t}) \;=\; \min_{t_i} \min_{\boldsymbol{t}\in\mathbb{S}_{i-1}^{t_i-1}} \ell(\boldsymbol{t}) + C(t_i,t) \;=$$

$$\min_{t_i} C(t_i,t) + \min_{\boldsymbol{t}\in\mathbb{S}_{i-1}^{t_i-1}} \ell(\boldsymbol{t}) \;=\; \min_{t_i} C(t_i,t) + \mathcal{V}_{i-1}^{t_i-1}$$

In this formula t_i ranges from $t-W$ to $t-w$. We have $T \times m$ values of $\mathcal{V}_m^T$ and calculating each takes at most $O(W)$ steps. The total time complexity is $O(TWm)$.

3. Posterior Marginal of Song Boundary

Fix a learning rate η, and fix T and M. Let

$$P(j,s) \;=\; \frac{\displaystyle\sum_{\boldsymbol{t}\in\mathbb{S}_m^T : t_j = s} e^{-\eta\,\ell(\boldsymbol{t})}}{\displaystyle\sum_{\boldsymbol{t}\in\mathbb{S}_m^T} e^{-\eta\,\ell(\boldsymbol{t})}}$$

That is, $P(j,s)$ is the "posterior probability" that song j starts at time s.

To compute $P(j,s)$, we need an extended notion of segmentation. We call $\boldsymbol{t}$ a $m/F:T$ segmentation if

$$F = t_1 < \ldots < t_m < t_{m+1} = T+1.$$

Let $\mathbb{S}_m^{F:T}$ be the set of all $m/F-T$-segmentations. We have

$$\sum_{\boldsymbol{t}\in\mathbb{S}_m^T : t_j = s} e^{-\eta\,\ell(\boldsymbol{t})} \;=\; \sum_{\boldsymbol{t}\in\mathbb{S}_{j-1}^{s-1}, \boldsymbol{t}'\in\mathbb{S}_{m-j+1}^{s:T}} e^{-\eta(\ell(\boldsymbol{t})+\ell(\boldsymbol{t}'))} \;=\; \left(\sum_{\boldsymbol{t}\in\mathbb{S}_{j-1}^{s-1}} e^{-\eta\,\ell(\boldsymbol{t})}\right)\left(\sum_{\boldsymbol{t}\in\mathbb{S}_{m-j+1}^{s:T}} e^{-\eta\,\ell(\boldsymbol{t})}\right)$$

which upon abbreviating

$$\mathcal{H}_m^t \;=\; \sum_{\boldsymbol{t}\in\mathbb{S}_m^t} e^{-\eta\,\ell(\boldsymbol{t})} \qquad\qquad \mathcal{T}_m^f \;=\; \sum_{\boldsymbol{t}\in\mathbb{S}_m^{f:T}} e^{-\eta\,\ell(\boldsymbol{t})}$$

means that we can write

$$P(j,s) \;=\; \frac{\mathcal{H}_{j-1}^{s-1}\cdot\mathcal{T}_{m-j+1}^{s}}{\mathcal{H}_m^T}.$$

So it suffices to compute $\mathcal{H}_m^t$ and $\mathcal{T}_m^t$ for all relevant t and m. We use

$$\mathcal{H}_1^t \;=\; e^{-\eta C(1,t)} \qquad\qquad \mathcal{T}_1^f \;=\; e^{-\eta C(f,T-f+1)}$$

and for $m \geq 2$

$$\begin{aligned}
\mathcal{H}_m^t \;&=\; \sum_{t_m}\sum_{\boldsymbol{t}\in\mathbb{S}_{m-1}^{t_m-1}} e^{-\eta(\ell(\boldsymbol{t})+C(t_m,t-t_m+1))} & \mathcal{T}_m^f \;&=\; \sum_{t_2}\sum_{\boldsymbol{t}\in\mathbb{S}_{m-1}^{t_2:T}} e^{-\eta(C(f,t_2-f)+\ell(\boldsymbol{t}))}\\
&=\; \sum_{t_m} e^{-\eta C(t_m,t-t_m+1)} \sum_{\boldsymbol{t}\in\mathbb{S}_{m-1}^{t_m-1}} e^{-\eta\,\ell(\boldsymbol{t})} & &=\; \sum_{t_2} e^{-\eta C(f,t_2-f)} \sum_{\boldsymbol{t}\in\mathbb{S}_{m-1}^{t_2:T}} e^{-\eta\,\ell(\boldsymbol{t})}\\
&=\; \sum_{t_m} e^{-\eta C(t_m,t-t_m+1)}\mathcal{H}_{m-1}^{t_m-1} & &=\; \sum_{t_2} e^{-\eta C(f,t_2-f)}\mathcal{T}_{m-1}^{t_2}
\end{aligned}$$

4. Posterior Marginal of Song Position

Fix a learning rate η, and fix T and M. Let

$$P(j,s,f) \;=\; \frac{\displaystyle\sum_{\boldsymbol{t}\in\mathbb{S}_m^T:t_j=s\wedge t_{j+1}-1=f} e^{-\eta\,\ell(\boldsymbol{t})}}{\displaystyle\sum_{\boldsymbol{t}\in\mathbb{S}_m^T} e^{-\eta\,\ell(\boldsymbol{t})}}$$

That is, $P(j,s,f)$ is the "posterior probability" that song j starts at time s and finishes at time f. In the same vein as the last section, we now get

$$P(j,s,f) \;=\; \frac{\mathcal{H}_{j-1}^{s-1}\cdot e^{-\eta C(s,f-s+1)}\cdot\mathcal{T}_{m-j}^{f+1}}{\mathcal{H}_m^T}.$$