Unsupervised Harmony Segmentation and Classification as an Optimization Problem

In order to harmonically cluster a midi file, we will make the assumption that the harmonic class of a "harmonically unitary segment" (contains only one class) is the same as another such segment if and only if it has a similar weighted combination of pitches within it. Each harmonically unitary segment is thus embedded in 12-dimensional space corresponding to the 12 pitch classes.

Now we split the problem into 2 sections. First we must create ideal harmonically unitary segments, as represented by chord changes across a song. Next we must cluster these and figure out the ideal harmonic classes. In both cases we are optimizing for cluster separation between harmonic classes. For simplicity let us assume a fixed amount of harmonic segments across the piece. This way, in the first problems we adjust the 12-dimensional space, while in the second part we are adjusting where our cluster boundaries fall.

Our goal is to maximize cluster separation. In order to do this we will break this up into 2 metrics: cluster center separation, and cluster tightness.

First, let us take the original midi file and deconstruct our input into pitch classes. For each pitch class we will have function $f_i(t)$, representing how prominent a pitch class is at time t. For simplicity we can take $f_i(t) = 1$ if pitch i appears at time t, and $f_i(t) = 0$ otherwise. In order to make this differentiable, we can combine sigmoid functions for this effect. For example, to signify note i is being played from time t to time s, and not otherwise:

$$f_i(x) = \frac{1}{1 + e^{\alpha(i-x)}} + \frac{1}{1 + e^{-\alpha(j-x)}} + 1$$

Where $\alpha > 0$ is the tightness of the sigmoid (high values should be used), and $i < j$.

This gives us, $f(t) \in \mathbb{R}^{12}$.

For each harmony, we will establish a cluster (assume 24 main harmony classes across one composer's body of work). Given cluster centers we will use least squared distance to calculate a metric for cluster center separation:

$$C(\{c_j\}_{24}) = \sum_{s,t} ||c_s - c_t||_2^2$$

Now we will evaluate a way to determine cluster tightness:

$$T(\{t_i\}_{100}, \{c_j\}_{24}) = \sum_{j} \sum_{t_i \in \mathcal{N}(j)} ||f(t_i) - c_j||_2^2$$

In order to formulate this as a differentiable function we will create, with a hyper-parameter σ:

$$\widetilde{T}(\{t_i\}_{100}, \{c_j\}_{24}) = \sum_{i,j} \exp(-\frac{1}{\sigma}||f(t_i) - c_j||_2^2)||f(t_i) - c_j||_2^2$$

This gives us the optimization problem:

$$min_{\{t_i\}_{100}, \{c_j\}_{24}} \left\{ -C(\{c_j\}_{24}) + \widetilde{T}(\{t_i\}_{100}, \{c_j\}_{24}) \right\}$$