标题：HMM Based Voice Separation of MIDI Performance

主要研究了使用 HMM 来将 midi 的复调变为单声。

优点：

- 可以使得同一音频中的音符略微重合，从而不需要对 midi 进行预处理，如数字化等；
- 可以对直播进行转换，也就是不需要输入全部 midi，能增量转换，因为该模型不需要设置一些诸如音频属性常量，也不需要完全分离所有音符。

建模过程：

由于音频为增量转换，故对于每一个音符，其长度定义为结束时的毫秒与开始时的毫秒之差：

$$\mathrm{Dur}(n) = \mathrm{Off}(n) - \mathrm{On}(n).$$

对于 HMM 里面的每一个状态，都由一系列单音 Vn 代表，其中每个单音 V 里面有诺干个音符序列。

对于相邻的两个音符，可以允许其略微重合，但必须满足两个条件，一个是前一个音符的结尾必须在后一个音符之前，另一个是前后音符重合的长度必须小于前面一个音符长度的一半。

$$\mathrm{Off}(n_i) - \mathrm{On}(n_{i+1}) \leq \frac{\mathrm{Dur}(n_i)}{2},$$

$$\mathrm{Off}(n_i) < \mathrm{Off}(n_{i+1}).$$

每个单音 V 还有音高 pitch 属性：

$$\mathrm{Pitch}(V) = \frac{\sum_{i=0}^{\min(l,|V|)} (2^i * \mathrm{Num}(n_{|V|-i}))}{\sum_{i=0}^{\min(l,|V|)} 2^i}.$$

每一个状态输出一组 midi 音符，这个音符来自于状态中的 所有单音 V 且满足如下条件：

$$\mathrm{On}(n) = \mathrm{Max}(\mathrm{On}(n')), \quad \forall n' \in V, \forall V \in S.$$

也就是最后一个音。

HMM 中状态由 S1 转换到 S2 转移条件是：S2 与 S1 的差集必须是 S2 的 output，也就是上面的的狮子；S2 与 S1 的并集必须与 S1 的出现顺序一样。

这个状态转换函数与 S、N、W 有关，S 即出事双胎，N 是 output 出的音符列表，W 是一个整数序列，每个值对应着原始 S 中的 V 序列中第几个音符需要被提取。

最后，概率由下式定义：

$$\mathrm{P}(T_{S,N,W}) = \prod_{0<i<|N|} \mathrm{P}(S, n_i, w_i) * \mathrm{order}(S, n_i, w_i). \quad (6)$$

最终示意大致如下：

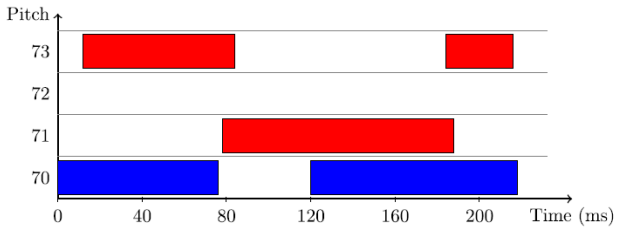

Fig. 2. An example of the notes that might be found in a MIDI file. Here, each note is colour-coded based on the voice to which our HMM would assign it.

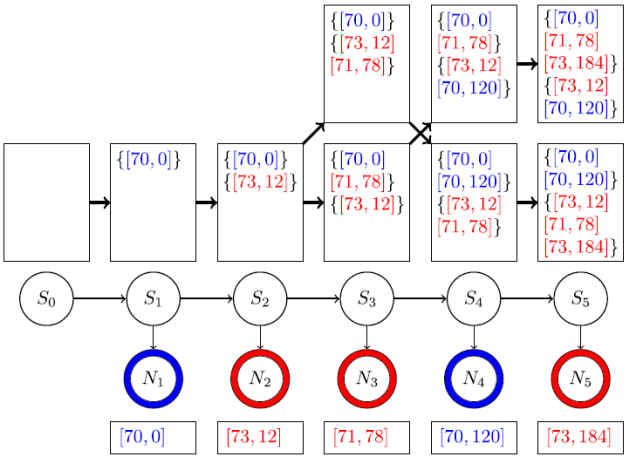

Fig. 3. An example of our model being run on the MIDI notes from Figure 2 with a beam size of 2. Each observed note set's border, and each note, is colour-coded based on the voice to which it is finally assigned. The two most likely state hypotheses at each step are listed in the large rectangles above the state diagram, with the more likely

这篇文章没看懂，但感觉似乎用处不是很大。