ECS7006 Music Informatics

Tutorial - Audio Matching & Cover Song Detection

1. For each of the below pairs of songs:

   - Which musical attributes are different?
   - What features would you extract in order to correctly detect cover versions?

Bob Dylan Knockin on Heavens Door 🔊	Avril Lavigne Knockin on Heavens Door 🔊
Metallica Enter Sandman 🔊	Apocalyptica Enter Sandman 🔊
Pink Floyd Wish You Were Here 🔊	Wyclef Jean Wish You Were Here 🔊

   Solution:

   - Knockin on Heavens Door: differences include (but are not limited to) key transposition and instrumentation; and vocals starting at different timings; the second version is also slightly faster and the rhythmic activity is more prominent due to guitar strums. A beat-synchronous cyclic chroma or cyclic CENS feature could work in this case.
   - Enter Sandman: differences include the presence/absence of vocals and percussion; instrumentation is clearly different, although the key is the same (E minor). A beat-synchronous chroma or CENS feature could work in this case.
   - Wish You Were Here: differences include tempo, rhythm/meter, instrumentation, vocal techniques, and chord progressions. Source separation on the vocals could help prior to the use of melody or chroma features on the separated vocals for detecting cover versions.

2. Consider the case of developing a system for *cross-modal audio matching*, where the user would select an image segment of a scanned and digitised music score and the system would find all corresponding audio fragments of similar music content. What would be the main building blocks of such a system?

   Solution:

   The system could have the following building blocks:

   - A component for performing optical music recognition (OMR) on the image segment selected by the user (this process could also be done offline). This component would take as input an image segment and would return a music score in machine-readable format (e.g. musicXML, MEI, MIDI...).

- A component for synthesizing the estimated machine-readable score into audio. Depending on the instruments present in the original score, some consideration on soundfonts and instrument identities to be used during synthesis.
- A final component would perform audio matching between the synthesized estimated score and the database recordings, similar to that described in the week 10 lecture / section 7.2 in the FMP book.

3. Let $c \in \mathbb{R}_{\geq 0}$ be a local cost measure defined by $c(x, y) = |x - y|$ for $x, y \in \mathbb{R}$. Given sequences $X = (x_1, \ldots, x_N) = (1, 2, 3)$ of length $N = 3$ and $Y = (y_1, \ldots, y_M) = (0, 3, 4, 1, 3, 3, 5)$ of length $M = 7$, compute the matching function $\Delta_{diag} : [0 : M - N]$ and the resulting best match.

   Solution:
   We compute the cost matrix $\mathbf{C}$ using the formula $\mathbf{C}(n, m) = c(x_n, y_m)$:

3	3	0	1	2	0	0	2
2	2	1	2	1	1	1	3
1	1	2	3	0	2	2	4
	0	3	4	1	3	3	5

   Then, for $m = [0, \ldots, 4]$, $\Delta_{diag}(m) = \frac{1}{N} \sum_{n=1}^{N} c(x_n, y_{n+m}) = \frac{1}{3}(3, 6, 4, 1, 5)$.
   The index m^* that minimises the above function is $m^* = 3$. The best match is $Y(1 + m^* : N + m^*)$ = Y(4:6) = (1,3,3).

4. Using the cost measure c, and sequences X and Y from exercise 3, compute the DTW-based matching function $\Delta_{DTW}[1 : M]$ using the step size $\Sigma = \{(1, 0), (0, 1), (1, 1)\}$.

   Solution:
   First, we need to compute the $N \times M$ accumulated cost matrix $\mathbf{D}$.

   The first column is initialised by: $\mathbf{D}(n, 1) = \sum_{k=1}^{n} \mathbf{C}(k, 1)$; and the first row is initialised by: $\mathbf{D}(1, m) = \mathbf{C}(1, m)$.

   Then, we use the following recursion to fill in the remaining values of $\mathbf{D}$:

$$\mathbf{D}(n, m) = \mathbf{C}(n, m) + \min \begin{cases} \mathbf{D}(n-1, m-1) \\ \mathbf{D}(n-1, m) \\ \mathbf{D}(n, m-1) \end{cases}$$

   $\mathbf{D} =$

3	6	2	3	3	1	1	3
2	3	2	4	1	1	2	5
1	1	2	3	0	2	2	4
	0	3	4	1	3	3	5

   The matching function Δ_{DTW} is given by: $\Delta_{\mathrm{DTW}}(m) = \frac{1}{N} \mathbf{D}(N, m) = \frac{1}{3}(6, 2, 3, 3, 1, 1, 3)$.