Structure, repetition and segmentation

Juan Pablo Bello
MPATE-GE 2623 Music Information Retrieval
New York University
Identifying structure from audio

- Arthur G. Lintgen: able to identify unlabeled recorded orchestral works by observing the spacing and patterns of grooves in an LP

- Inspired J. Foote (ISMIR, 2000) to develop a MIR system based on structural similarity
Musical Form

- Units can be assigned letters (A, B, C) or functional names (intro, verse, chorus, bridge, etc).

- Strophic: repeats the same section, e.g. AA...

- Binary: alternates two sections, which are often repeated, e.g. ABAB or AABB

- Ternary: third section is often a variation of the first, e.g. AABA, AABA’, AA’BA’

- Arch: symmetric, repetition of sections around a center, e.g. ABCBA

- Rondo: main theme is alternated with sub-themes, e.g. ABACADA.....

- Variations: theme plus variations, e.g. AA\textsuperscript{i}A\textsuperscript{ii}AA\textsuperscript{iii}

- Sonata: complex developmental form including the exposition, development and recapitulation of a given theme(s).
Repetition

• Musical form is often defined by the amount of repetition across sectional units.

• Repetition is central to music (in harmony, melody, rhythm, etc).

• Significant variations are often found between repeated parts.
The information necessary to characterize repetitions is encoded in the feature sequence:

(a)

(b)
Feature vectors

- Frame-level feature values represent a point (or vector) in the N-dimensional space defined by the different features (e.g. 12 chromas, 15 MFCCs, etc):
Feature vectors

• By calculating how (dis)similar these vectors (or groups of vectors) are we can characterize the structural organization of the feature sequence.
Self-similarity matrix

- For an N-long sequence of feature vectors
- Self-similarity matrix: N x N matrix of pairwise (dis)similarities between vectors
Self-similarity matrix

- Vertical and horizontal axes represent time
- Symmetric similarity function = symmetric matrix
- Main diagonal: closer/most similar values
- Similar subsequences (repetitions) -> diagonal stripes in the plot
Self-similarity matrix
Self-similarity matrix
Self-similarity matrix
Self-similarity matrix
Measuring similarity

• Let us consider two feature vectors, a and b, each representing a distinct segment of an audio signal:
Measuring similarity

• Euclidean distance: length of the straight line between points (vectors) in the feature space.

• In N-dimensional spaces, for two points $a$ and $b$, it is defined as:

$$d = \sqrt{\sum_{i=1}^{N} (a_i - b_i)^2}$$
Measuring similarity

• The Euclidean distance is a special case of the $L_p$ norm of the $N$-dimensional vector of differences:

$$d = \left( \sum_{i=1}^{N} |a_i - b_i|^p \right)^{\frac{1}{p}}$$

• Euclidean distance corresponds to the $L_2$-norm

$$\|x\|_2 = \left( \sum_{i=1}^{N} |x_i|^2 \right)^{1/2}$$

• Manhattan or taxicab distance -> $L_1$-norm

$$\|x\|_1 = \sum_{i=1}^{N} |x_i|$$

• Chebyshev distance -> $L_\infty$-norm

$$\|x\|_\infty = \max[|x_i|]$$
Measuring similarity

• Alternatively we can use the dot product between a and b:

\[ a \cdot b = \sum_{i=1}^{N} a_i b_i = ||a|| ||b|| \cos \theta \]

• Geometrically: product between the length of b (||b|| = ||b||_2) and the scalar projection of a into b; \( \theta \) is the angle between the vectors.
Measuring similarity

• The dot product will be large if both vectors are large and similarly oriented.

• We may want to make this operation independent of vector length, thus we normalize the dot product:

\[
\cos \theta = \frac{a \cdot b}{\|a\| \|b\|}
\]

• The resulting metric is known as the cosine distance.

• If \( a \) and \( b \) have zero mean \( \rightarrow \) the cosine distance measures also the correlation between the vectors.
Enhancing SSM

- Time delay embedding (Marwan, 2006): concatenating m neighboring vectors spaced by a delay $\tau$
Enhancing SSM

• Time delay embedding (Marwan, 2006): concatenating m neighboring vectors spaced by a delay $\tau$
Enhancing SSM

- Contextual similarity (Mueller & Kurth, 2007): use “tempo-warped” feature sequences to maximize local similarities (accounts for variable tempo).
Enhancing SSM

• Transposition invariance (Goto, 2003): similarity btw chromagram and all cyclically-shifted versions of itself; $SSM_t \rightarrow$ pointwise maximum btw SSMs

(a) (c) (e)

(b) (d) (f)

standard transposition invariant
SSM-based segmentation (Serra et al, 12)
SSM-based segmentation (Müller and Kurth, 07)

- Path extraction: Initialization

- (1) threshold SSM; (2) remove info on lower triangular part; (3) remove main diagonal and its neighborhood
SSM-based segmentation (Müller and Kurth, 07)

- Path extraction: Construction

- (1) choose max similarity value and create new path P; (2) extend P in all possible directions while SSM values are above a threshold
SSM-based segmentation (Müller and Kurth, 07)

- Path extraction: Removal + iteration

- (1) we remove all monotonically decreasing values in horizontal/vertical “rays” from each cell in path P; (2) we repeat construction-removal steps until no SSM value is above a threshold.
SSM-based segmentation (Müller and Kurth, 07)

- Clustering:
  - (1) Each path describes a pairs of audio segments
  - (2) Transitivity: for all segments, group all overlapping segments in the path list
  - (3) Merge groups corresponding to segments in the same path
  - (4) discard redundant groups
References


