Analyzing structure: segmenting musical audio

Musical form (1)

- Can refer to the type of composition (as in multi-movement form), e.g. symphony, concerto, etc
- Of more relevance to this class, it refers to the structure of a given piece (as in single-movement form), e.g. strophic, binary, sonata, fugue
- We can differentiate between sectional and developmental form
- “In a sectional form, the piece is built by combining small clear-cut units, sort of like stacking legos” (DeLone, 1975). This is the case with most popular music.
- In developmental form the piece is built from a number of evolving presentations and combinations of small musical units (e.g. motifs, themes)
- Music with continuous non-sectional, non-repetitive form is called *through composed*

Musical form (2)

- Each unit can be labeled with a letter (e.g. A, B, C) or a generic name (e.g. intro, verse, chorus, bridge, interlude, coda, etc)
- Strophic form: repeats the same tune in different verses, e.g. AA…
- Binary form: alternates two sections, which are often repeated, e.g. ABAB or AABB
- Ternary form: has three parts, third section is often a recap or a variation of the first one, e.g. AABA, AABA', AA'BA'
- Arch form: it is a symmetric form, based on the repetition of sections around a center, e.g. ABCBA
- Rondo: a main theme is alternated with sub-themes, but it always comes back (returns), e.g. ABACADA…..
- Variations: theme plus variations, e.g. $AA^{i}A^{ii}AA^{iii}$
- Sonata: more complex form showing intro, exposition, development, recapitulation, coda.

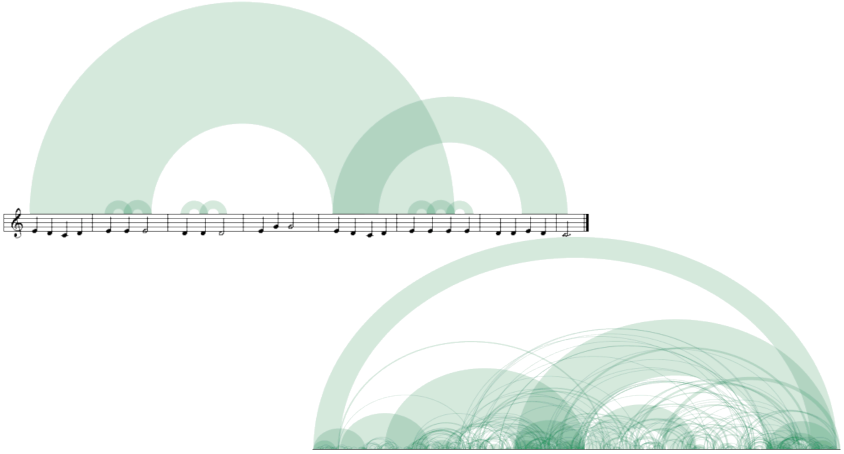
Musical form (3)

- What is at the core of structure analysis is the idea of repetition.
- Music tends to be highly repetitive, thus by identifying those repetitions we can characterize long-term structure
- But, what type of repetitions? Melodic? Harmonic? Rhythmic?
- Mr. Arthur G. Lintgen is able to identify unlabeled recorded orchestral works by observing the patterns of grooves in an LP

Figure 1. Arthur G. Lintgen identifying a phonograph record by examining the grooves

Visualizing structure (1)

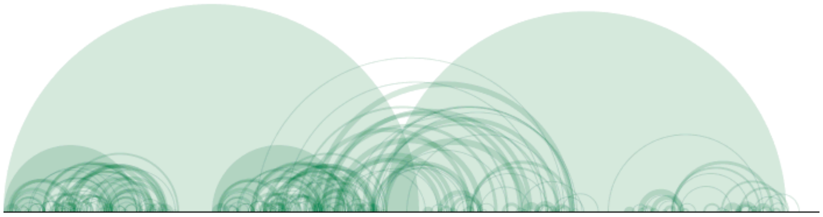
- We need representations of musical features (e.g. pitch) where these repetitions can be characterized



- Martin Wattenberg: http://www.bewitched.com/match/music.html

Visualizing structure (2)

- Bach's Minuet in G Major
- Binary form: AABB
- A and B overlap and are related (thin arcs)

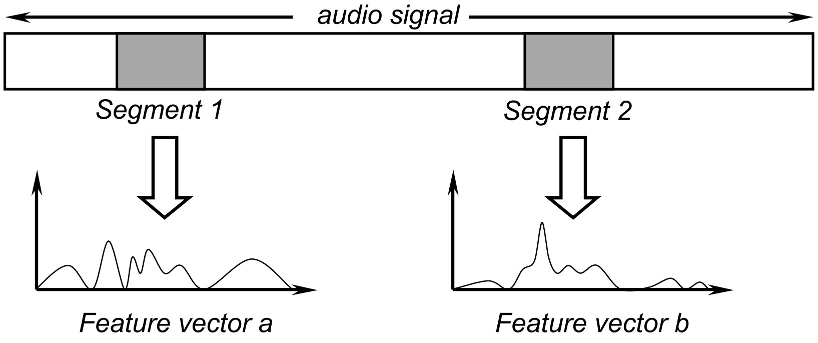


Audio-based analysis

- This representation relies on high-level musical info (pitch) and can only characterize repetitions of the exact same events.
- However, from audio we cannot obtain such high-level info without tolerating error
- The only reliable calculations are of low-level features such as spectral features (centroid, spread), MFCCs, LPCs, chroma, etc
- Thus we need to be able to find low-level feature sets that are able to characterize “repetitions”, and we need to be able to identify soft (approximate) repetitions.

Feature sets

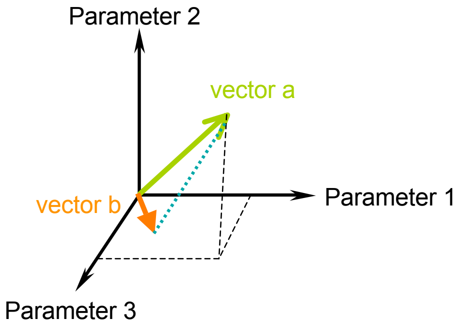
- Let us consider two feature vectors, a and b, each representing a distinct segment of an audio signal:



- We can calculate how different these two vectors are.

Feature sets

- Each feature set represents a vector in the Euclidean space defined by the different features (e.g. 12-D chroma vectors, 15-D MFCCs, centroid):

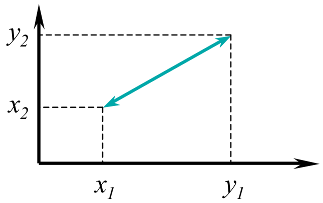


- These vectors have lengths (magnitudes) and angles, and we can calculate the distance between them

Distance metrics (1)

- In this space the simplest distance we can calculate between two points is that described by a straight line between them.
- That distance is known as an Euclidean distance and is defined (in 2D) as:

$$d = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}$$



- In n-dimensional (Euclidean) spaces, for two points a and b, it is defined as:

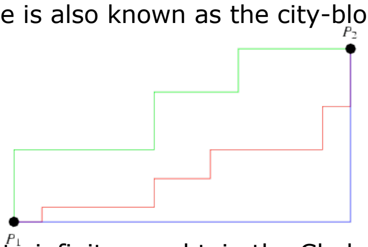
$$d = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Distance metrics (2)

- In the Euclidean space, other distances (related to norms) can also be used
- The general case, the L_p-distance in n dimensions, is defined as:

$$L_p - distance = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

- It can be seen that the Euclidean distance is the L_2-distance.
- The L_1-distance is also known as the city-block or Manhattan distance.



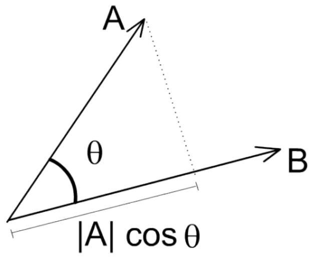
- When p tends to infinity we obtain the Chebyshev distance

Distance metrics (3)

- Alternatively we can use the dot product between two n-dimensional vectors (a and b) which is defined as:

$$a \cdot b = \sum_{i=1}^{n} a_i b_i = |b||a| \cos\theta$$

- Geometrically it can be interpreted as the product between the length of B ($||B|| = B \cdot B = (\sum(b_i)^2)^{1/2}$) and the scalar projection of A into B:



where θ is the angle between the two vectors

Distance metrics (4)

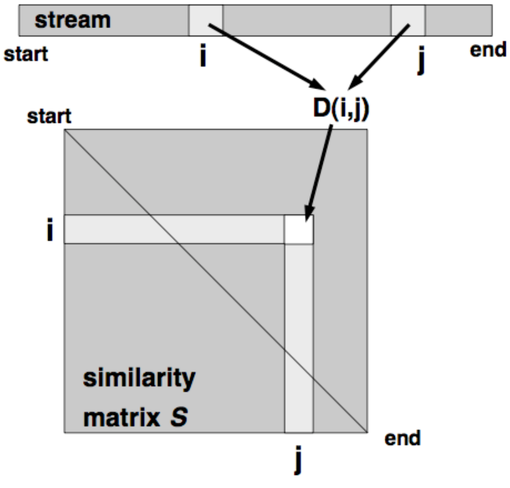
- The dot product will be large if both vectors are large and similarly oriented
- However, in some cases, we may want to make this operation independent of magnitude (of the vector lengths), thus we normalize the dot product such that:

$$\cos\theta = \frac{a \cdot b}{|b||a|}$$

- The resulting metric, the cosine of the angle between both vectors is know as the cosine distance
- If a and b have zero mean (which can be done by subtracting the mean from all values in the vector) then the cosine distance measures also the correlation between vectors.
- There are many other distance/correlation metrics: Mahalanobis, Earth-Mover’s distance, KL divergence, etc

Self-similarity matrix (1)

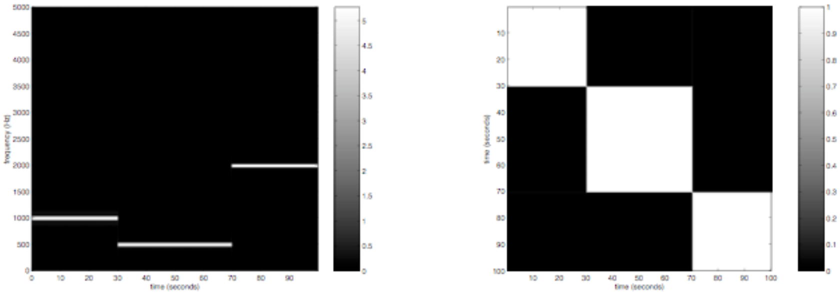
- We can recursively calculate the distance that separates the frame-by-frame feature vectors of an audio stream
- The resulting representation is known as a self-similarity measure, and depending on the actual metric it measures the (di)similarity between vectors.
- This is suited to represent repetitions, thus long-term structure in music



Similarity matrix calculation

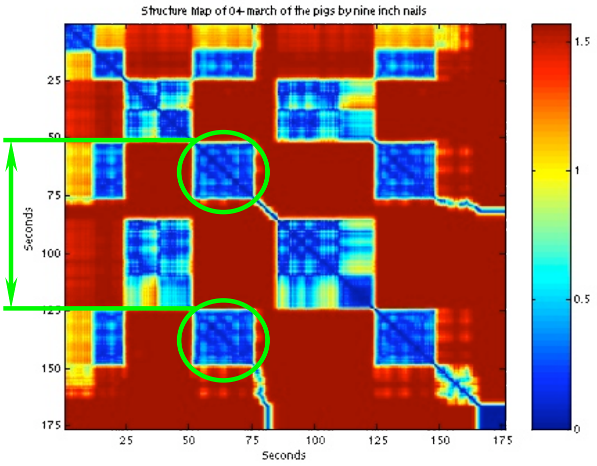
Self-similarity matrix (2)

- What to look for on a self-similarity matrix?
- Synthetic example (3 pure tones on the frequency domain)
- The main diagonal of the matrix is always the area of strongest self-similarity, corresponding to the autocorrelation of each vector



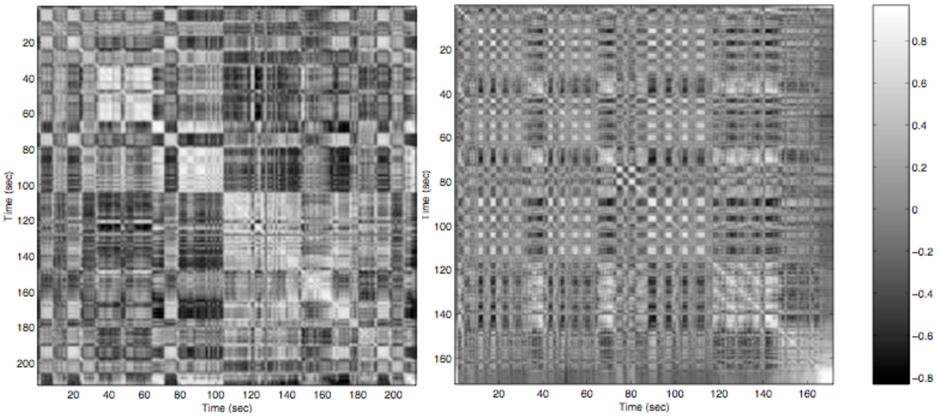
Self-similarity matrix (3)

- Other diagonals (and bright-colored blocks) are telling about possible repetitions and their location



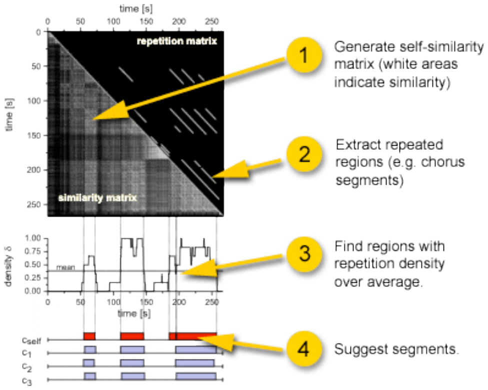
Self-similarity matrix (4)

- Some examples (Cooper and Foote, 2002): Vivaldi’s Spring and The Magical Mystery Tour by The Beatles.
- Features: MFCCs ; metric: cosine distance



Self-similarity matrix (5)

- The process of extracting structural information (identifying segment boundaries) from a self-similarity matrix is not trivial



1 Generate self-similarity matrix (white areas indicate similarity)

2 Extract repeated regions (e.g. chorus segments)

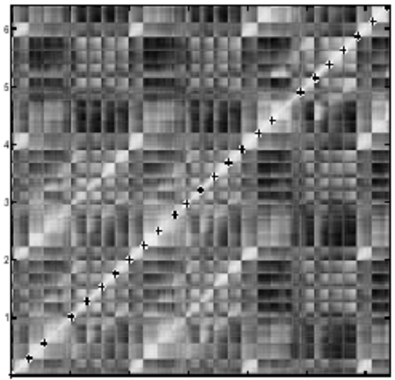
3 Find regions with repetition density over average.

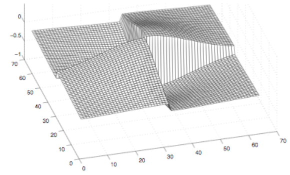
4 Suggest segments.

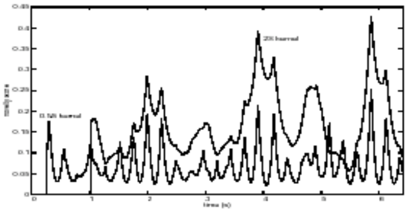
- There is no standard approach. Strategies are driven by application and feature set (see following examples)

Novelty function

- Foote (2000) uses the cosine distance on STFT coefficients
- A checkered kernel is thus passed through the diagonal of the matrix to quantify changes in a novelty function
- Small kernels work for onset detection, while large kernels characterize longer segments on the signal

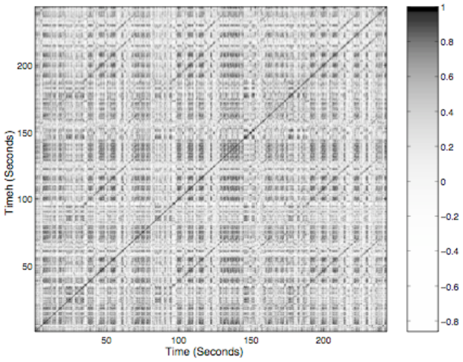


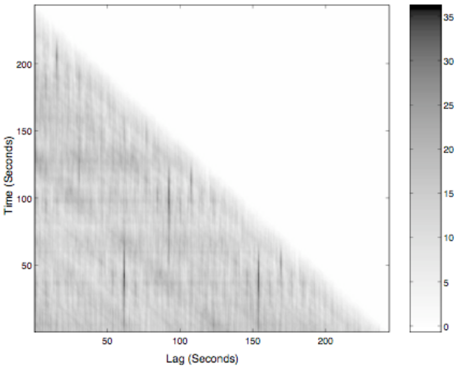




Thumbnailing (Summarization)

- Bartsch and Wakefield (2001) use beat-synchronous chroma vectors and cosine distance to generate S.
- Then they filter along the diagonals of the matrix with a moving-average filter to identify regions of extended similarity (characterized by lines of constant lag in the direction of the columns)
- Thumbnails are selected by locating the area of similarity that carries the most energy


Timeh (Seconds)
Time (Seconds)


Time (Seconds)
Lag (Seconds)