The Beatles Genome Project: Cluster Analysis and Visualization of Popular Music

Douglas J. Mason[*]
Harvard University

Abstract

We present a database system for storing and retrieving abstracted musical information as well as visualizations for interpreting this information. We then apply hierarchical cluster analysis to show statistical phenomena occuring in a corpus of popular songs written by the Beatles. We find that chords and melodic rhythms fall into distinct clusters which distinguish each song, indicating possible principles of composition.

Index Terms: H.3.3 [Clustering]: Information Search and Retrieval—Information Storage and RetrievalInformation Systems; H.5.5 [Methodologies and techniques]: Sound and Music Computing—Information Interfaces and PresentationInfromation Systems

1 Introduction

Sheet music notation serves performers rather than analysts. As a result, rhythmic values and absolute chromatic pitches take precedence over the relationship of melody to harmony, rhythm, or key. Recent developments in musical information storage and retrieval attempt to ameliorate the situation by encasing musical objects in abstractions which can then be manipulated and rendered at-will on the computer. With the recent introduction MIT's Music21 Python package[1], musical objects can now be married to state-of-the-art techniques in computational musicology that allow for the automation of musical analyses. In this poster, we use Music21 to drive visualizations that emphasize statistical aspects of the musical data. We combine two such visualizations: the SongMap – which succinctly normalizes individual songs and emphasizes structural components to facilitate comparisons – and the hierarchical clustering heatmap – which can indicate strong statistical threads throughout the entire corpus.

2 Our Approach

Unlike audio analysis which tends to emphasize production, timbre, and beat, our analysis uses a single melodic line written in standard music notation, along with chords (the co-occurence of multiple notes in the background) which sound simultaneously. Our software front-end divides the task of musical input into individual phrases that can be typed using an easy-to-understand textual musical language called TinyNotation, developed at MIT and enhanced for our purposes. Phrases are linked together into larger units (the verse, chorus, and bridge), which are then strung together to form the entire song data-stream (verse-verse-bridge-verse).

Once a song has been stored in this format, it can be retrieved and processed using Music21 to generate visualizations called SongMaps, which are designed to emphasize elements of interest to the analyst. For instance, SongMaps are normalized to the song's tonal center and emphasize the distribution of stable/unstable tones in the melody, while rhythmic syncopation is identified along the bottom to indicate structural choices. The length and placement of musical phrases against the bar lines, along with their harmonies, take precedence by forming the SongMap's foundation. Each visualization fits on a single page to facilitate comparative analysis between songs and song sections.

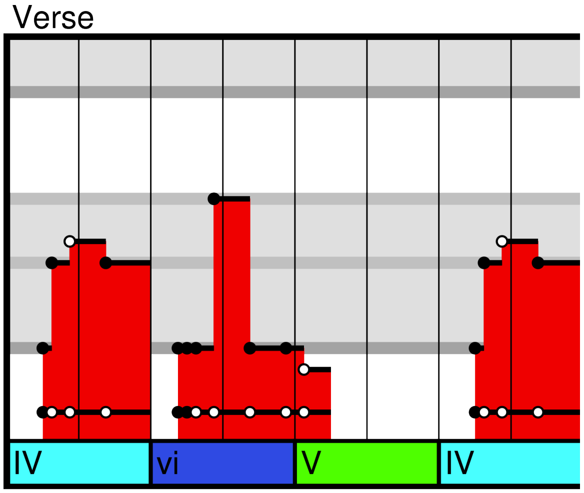

Figure 1: Portion of a SongMap for the self-composed "I Remember", showing harmony (colored boxes and labels along the bottom), melody (black/white circles on top of red boxes, whose height indicates pitch), rhythm (black/white circles along bottom), and stability tones (grey boxes in background). Time moves from left to right and white dots indicate unstable pitches or syncopated rhythms. Each SongMap fits into the same rectangular region and is normalized to the tonal center of the song, allowing for gestalt comparison between songs.

3 Results

SongMaps make it possible to cultivate a gestalt impression of a corpus of music by presenting songs side-by-side in a normalized fashion. In the case of this poster, we have focused on the Rolling Stone list of the 100 Greatest Beatles Song[2], which have all been entered into the database and cross-checked against The Beatles Fake Book[3].

We quickly noticed the inclusion of unusual chords in The Beatles corpus (this has also been observed in other corpuses of rock music, see de Clercq and Temperley[4]). In common practice classical music, the vast majority of harmonies fall on the diatonic scale, and can be written in Roman numeral notation as I, ii, iii, IV, V and vi (the number indicates the scale degree of the chord root, normalized by the tonal center; upper-case indicates major tonality, and lower-case indicates minor tonality). For the Beatles corpus, chords like bVII, bIII, and bVI, which are borrowed from other scales, are surprisingly common, although the corpus is still dominated by diatonic harmonies.

To examine the presence of these unusual chords, we describe each song in a feature vector space that links each dimension to an

[*] e-mail: douglasmaosn@gmail.com

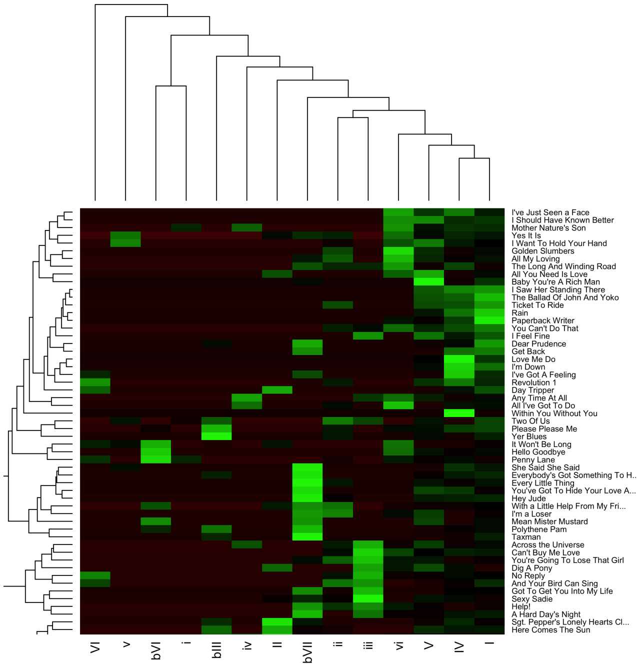

Figure 2: We perform cluster analysis on chords using tf*idf measures. Above, a snippet of the chord-cluster heatmap showing the strong grouping of songs based on the presence of non-diatonic chords (all chords except I, ii, iii, IV, V, and vi). Green indicates the unusual presence of a chord in a song, while red indicates its notable absence. Obvious clusters can be found, such as the bright marking for the bVII chord in the middle of the plot.

individual chord in relation to the tonal center. The value of a song in each dimension corresponds to the percentage of the song which plays that chord. We then normalize our feature space by applying tf*idf measures[5] to indicate the level of surprise for each chord. After applying hierarchical cluster analysis[6], we find that most songs in the corpus (∼70%) strongly emphasize unusual chords, and among those, the majority (∼80%) exhibit the strong presence of only one. Moreover, song clusters span many albums and almost never include two songs from the same album. Both of these results suggest that the Beatles were deliberate in limiting the harmonic palette of each song and careful not to repeat that palette in the same album.

Because of the richness of our dataset, we are able to examine elements of melody as well as harmony. In the SongMaps, the onset of musical phrases and their relation to the beats of the measure are strongly emphasized because it has been argued that this is a key element in a song's composition. To examine this assumption, we performed hierarchical cluster analysis on beat-onset of musical phrases. Because division between musical phrases is a subjective measure, we elected to define a feature space in which each dimension corresponds to a beat (or half-beat) of the measure, and the value corresponds to the percentage of melody notes which fall on that beat. These values are weighted by the length of time to the previous note to emphasize notes that start musical phrases. Then a tf*idf measure is applied to emphasize the element of surprise.

Our analysis indicates high-level clustering of songs with on-beat rhythms (with integer beat values) or off-beat rhythms (with integer-plus-1/2 beat values). We also find strong clustering at the lower level which emphasizes individual beat onsets. Each cluster includes songs from many albums, although due to the limited dimensionality of beat values (8 possible beats and off-beats compared to 24 possible chords) it is common for multiple songs in the same album to cluster together. Like the chordal analysis, however, almost all songs emphasize only one beat value, indicating that the beat-onset of musical phrases acts like a musical fingerprint.

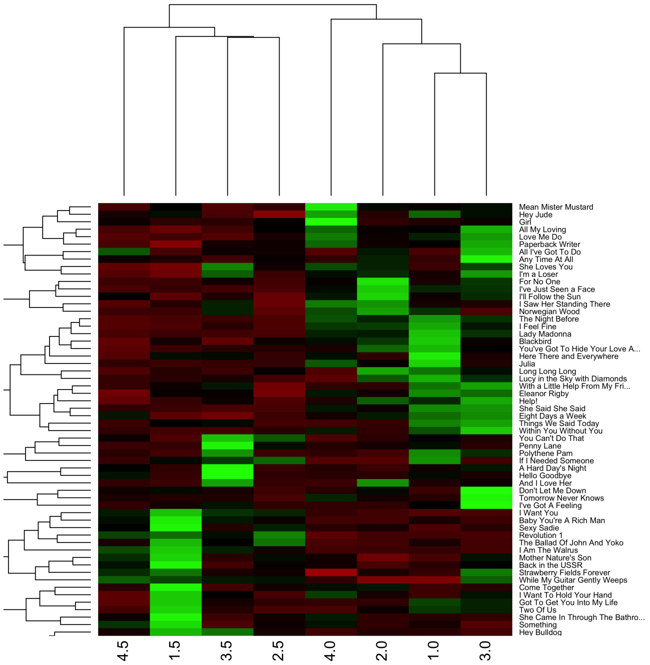

Figure 3: We also cluster songs using tf*idf measures on the beat of each note, weighted by the temporal distance to the note before. Above, a snippet of our analysis reveals strong clusters along each beat in green, suggesting that the beat on which melodic phrases start acts as a sonic fingerprint. Just as unusual chords tend to appear in isolation, we find that the majority of songs emphasize only one starting beat.

4 Conclusion

We have demonstrated the use of a large musical database which encodes abstractions of popular music, and derived visualizations to help interpret compositional principles from a corpus – in our case, 100 songs written by The Beatles. By applying hierarchical cluster analysis, we have produced robust statistical measures that identify distinguishing features of individual popular songs, such as the presence of an unusual chord or the tendency for musical phrases to start on the same beat of the measure. We find that our corpus organizes nicely along on-beat and off-beat rhythms.

Acknowledgements

The author wishes to thank Stan Swiniarski at Berklee College of Music and Michael Cuthbert at MIT

References

[1] Michael Scott Cuthbert and Christopher Ariza. music21: A toolkit for computer-aided musicology and symbolic music data. *Proceedings of the International Symposium on Music Information Retrieval 2010*, pages 637–42.

[2] The 100 greatest beatles songs. *Rolling Stone*, September 2011.

[3] *The Beatles Fake Book (C Edition)*. Hal Leonard Corporation, 1987.

[4] Trevor de Clercq and David Temperley. A corpus analysis of rock harmony. *Popular Music*, 30:47–70, 2011.

[5] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(11-21), 1972.

[6] M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press: New York., 1973.