Music Chord Progression Analysis

Final Project Report for “Probability and Statistics for Data Science” (ECE 225)

Hao-Wen Dong
hwdong@ucsd.edu

Chun-Jhen Lai
c8lai@eng.ucsd.edu

Introduction

We aim to analyze the statistical properties of chords and chord progressions in over ten thousand songs available on the HookTheory platform.[1] In particular, we are interested in analyzing the prior probabilities for different chords, the transition probabilities between chords and chords and the most frequent chord progressions. We are also interested in how these statistics differ from genre to genre with an eye to reveal some interesting trends on the usage of chords and chord progressions in different genres.

Data

HookTheory is a community-based platform for users to upload lead sheets (i.e., melody and chord progressions) for different songs. We use the version compiled and provided on a GitHub repository.[2] This dataset contains 18986 music segments from 11380 songs of 4956 artists.

Data preprocessing

The data is stored in JSON format. We first extract the metadata from each song. This includes key, speed (in bpm; beats per minutes), beats per measure and genre tags. We then look into the chords used in each song. We note that there are two versions of chord annotations, which are absolute chords (i.e., the actual chords played) and relative chords (i.e., the chords relative to the key signature)[3]. *In order to see statistics over different keys, we adopt the relative chord annotations.* In this way, this is similar to transposing each song into C key and conducting analysis on a collection of songs in C key. Moreover, there can be consecutive, duplicate chords in each chord sequence. Hence, we need to remove these duplicate chords from the chord sequence (e.g., CCFFGGAmAm → CFGAm) in order to acquire accurate statistics.

---

[1] https://www.hooktheory.com/
[2] https://github.com/wayne391/lead-sheet-dataset
[3] This is based on Roman numeral notation system used in musical analysis.

Metadata analysis

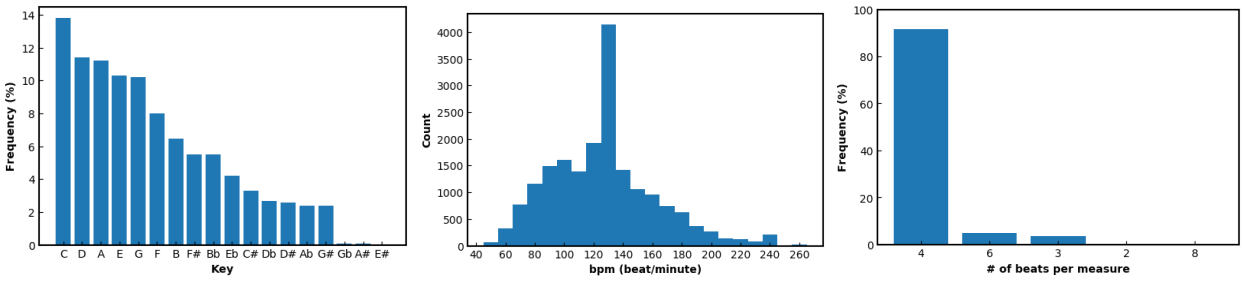
In order to gain an overview of the dataset, we first investigate some basic properties of the dataset. We plot the key distribution, the speed histogram and the distribution of beats per measure, as shown in Figure 1(a)-(c), respectively.

1. **Key distribution**—It is not surprising that C key is the most commonly used key, as there are no accidental marks (e.g., sharps and flats) in C key and it is the default key of the platform. Moreover, we can also see that G♭, A♯ and E♯ keys are the most rarely used keys. This is consistent to music theory. For example, E♯ key is actually a theoretical key signature.

2. **Speed**—The speed distribution is roughly bell-shaped with a spike at 128 bpm as it is the default value of bpm in the platform. Few songs have bpm lower than 50 or higher than 250.

3. **Beats per measure**—Four beats per measure represents over 90% of the dataset, which is followed by six and three beats per measure.

These metadata analysis results match our expectation and impression of real world music.



(a) key (b) speed (bpm) (c) beats per measure

Figure 1. Metadata analysis results.

Chord analysis

1. Chord quality

We first examine the chord quality distribution. As can be seen from Figure 2, major chords (‘M’) represent the majority (more than half), and about 35% are minor chords (‘m’), while major seventh chords (‘maj’), diminished chords (‘o’) and half-diminished chords (ø) are rarely used.

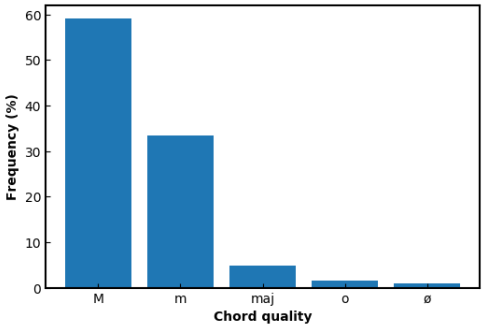


Figure 2. Chord quality distribution.

2. Chord prior probabilities (distribution)

We can from Figure 3 that C chord is the most frequently used. The distribution matches our impression of today's music. For example, common chords (in our impression) such as C, G, Cm, and F represent the most frequently used, while uncommon chords (in our impression) such as $D\sharp_{maj}$, B^{o}, $D^{\emptyset}$, $A\sharp$ and D^{o} are rarely used. Note that C, G and F represent the tonic, dominant and subdominant chords of C key, which are functionally important in music theory.

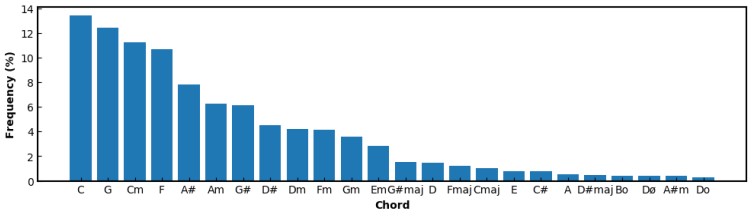
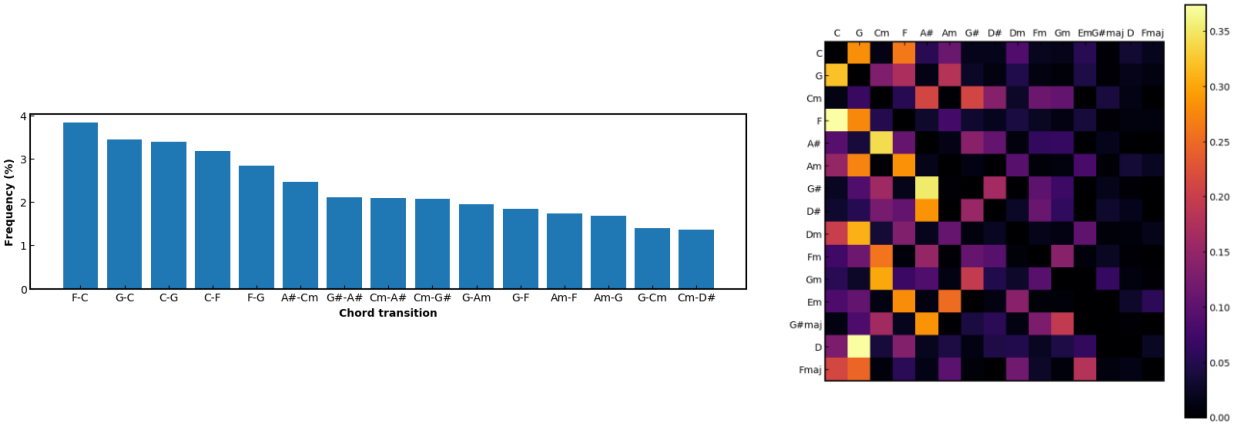


Figure 3. Chord prior probabilities (distribution) for the 24 most common chords.

3. Chord transition probabilities

We are interested the transition probabilities between different chords. First, we compute the frequencies of different chord transitions, as is shown in Figure 4(a). We can see that Transition between C, G and G, the three functionally important chords in music theory, are the most common. Moreover. we also visualize in Figure 4(b) the transition matrix between common chords. On one hand, we see that the transition matrix is asymmetric, which means that chord transitions is direction-sensitive. On the other hand, we also observe that the transition matrix is approximately a sparse matrix (i.e., many entries are close to zero). This can be an evidence of music rule as some chord transitions are theoretical unpleasant according to music theory.



(a) chord transition distribution (the most common 15)

(b) transition matrix for common chords (colors represent probabilities)

Figure 4. Chord transition probabilities.

4. Three-gram and four-gram analysis for chord progressions

Next, we conduct three-gram and four-gram analysis to see the most common chord progressions. As shown in Figure 5, the most common three-grams are F-C-G, C-F-C, F-G-C, and G♯-A♯-Cm, and the most common four-grams are C-G-Am-F, C-F-G-C, G-C-F-G, Cm-G♯-A♯-Cm and their cyclic permutations (i.e., C-G-Am-F and F-C-G-Am). In our impression, C-G-Am-F and C-F-G-C are indeed the two most frequently used chord progressions.

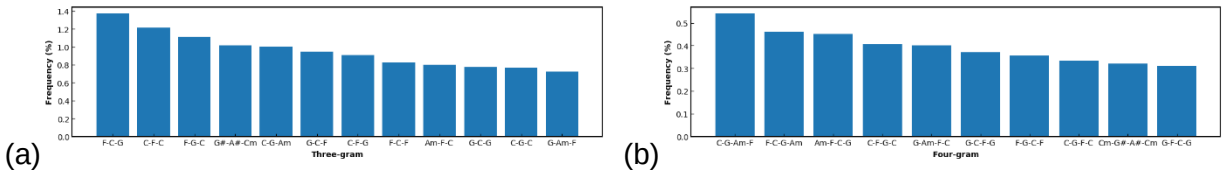


Figure 5. Distributions of (a) three-grams and (b) four-grams.

5. Chord and chord transition analysis per genre

We first compute the genre distribution, and we can see from Figure 6 that video game, pop, electronic, rock and soundtrack represent the majority. We then repeat the statistical analysis for each genre. As we can see from Figure 7 (a), pop, rock and alternative music use similar chords with similar frequencies. This indicates these three genres might be largely overlapped, which is generally true in today's music. A similar correlation is also observed between video game and soundtrack. Interestingly, minor chords are adopted more in R&B compared to that in other genres. What's more, from Figure 7 (b), we can see that the transition matrices for video game, J-pop and soundtrack are sparser, while R&B has a denser transition matrix. We interpret this as the fact that R&B music has higher perplexity and that the rule deployed by composer might be looser.

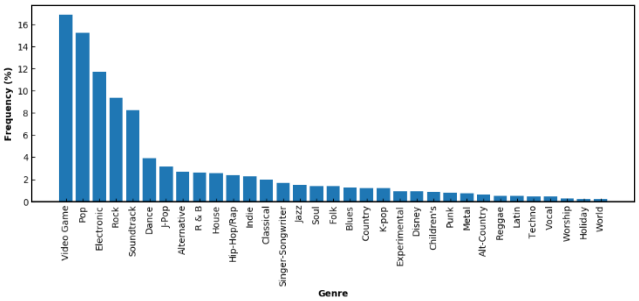
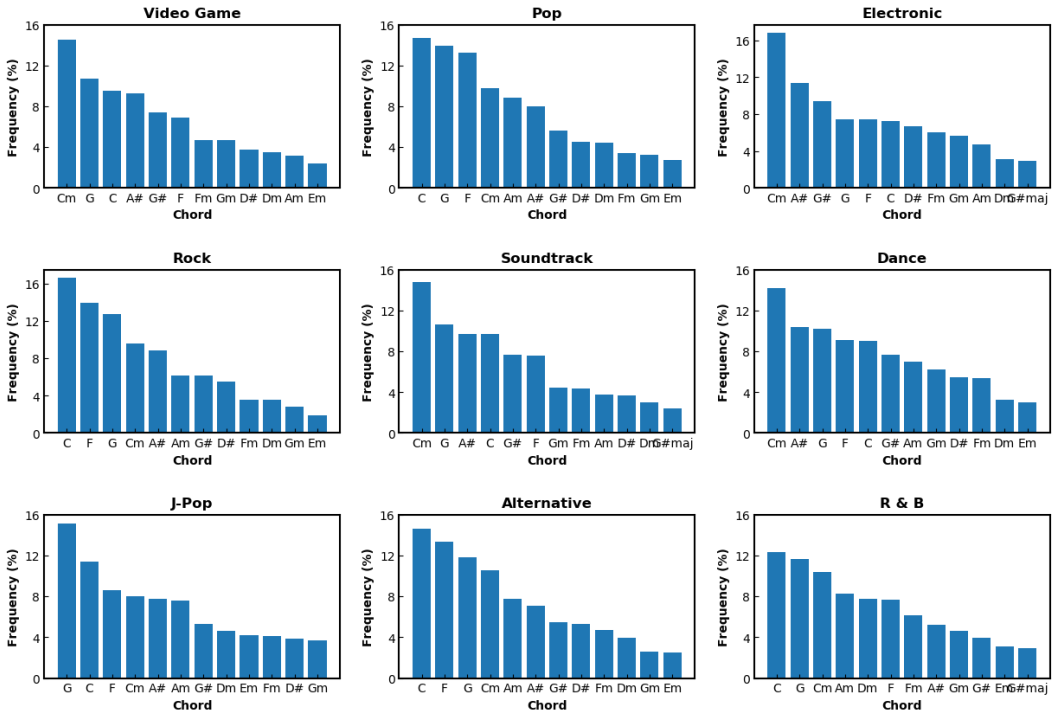


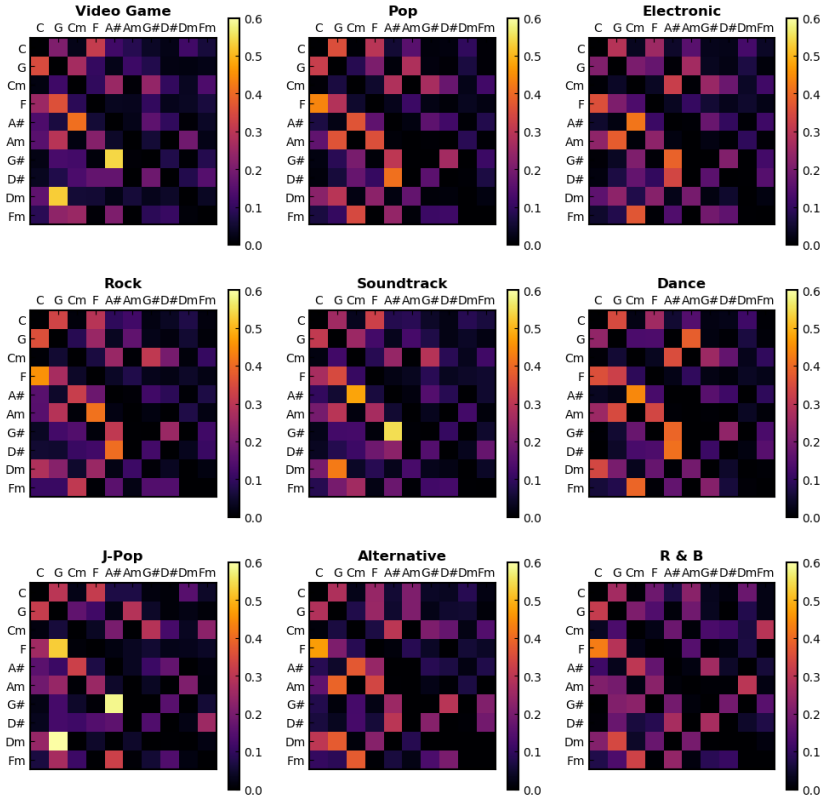
Figure 6. Genre distribution.

Conclusions

By conducting statistical analysis on this dataset, we gain some insights into the chords and chord progressions used in today's music. Most observed trends are consistent to our impression and understanding of music, while it is nice to adopt statistics to examine our thoughts. On the other hand, the difference in chord transition matrices for different genres is a surprising finding. From a viewpoint of information theory, we can use the entropy of transition matrix as a metric to quantify the perplexity of a genre, which might be an interesting future research problem.



(a) chord distribution per genre



(b) chord transition matrix per genre

Figure 7. Chord and chord transition analysis per genre.