Pitfalls and Windfalls in Corpus Studies of Pop/Rock Music

Introduction

Hello, my talk today is about corpus studies of pop/rock music—specifically, the benefits or "windfalls" of this type of work as well as some of the problems. I call these problems "pitfalls" because they are traps that we should do our best to recognize—either in our own work or in the work of others—and avoid whenever possible.

Background

[NEXT] At a basic level, a corpus study is any methodological investigation of some body of work, although in its modern meaning, it implies some sort of statistical analysis of encoded music using a computer. Corpus studies thus purport to give a more objective view of theoretical insights that were previously gleaned primarily through intuition alone. [NEXT] Corpus work has thus, not surprisingly, emerged as a promising subfield in music research, as shown by recent multiple-volume special issues.

As this type of work has become more commonplace in music overall, so too has the use of corpus methods to study pop/rock music. [NEXT] Existing corpora of pop/rock music include those created by a single author, such those by Jay Summach and David Tough. [NEXT] Pop/rock corpora also include those with multiple authors, such as the corpus by Ashley Burgoyne and others of top songs from the *Billboard* charts, as well as the corpus I created with David Temperley of the top songs reported by *Rolling Stone* magazine. [NEXT] Finally, some corpora of popular music have been created by computer algorithms, such as the "Million Song Dataset."

One of my central points today is that a corpus created by a single author or a computer algorithm has built-in shortcomings, and even a corpus created by multiple authors can fall short in similar ways. The primary problem derives from the subjective nature of music analysis, which can strongly shape the corpus and its results. The good news for music analysis, as I will argue, is that we can turn this pitfall into a windfall by using multiple independent annotators to assess subjectivity. In what follows, I discuss these issues via two main aspects of corpus work: 1) data collection and representation, and 2) data analysis and interpretation.

Data Collection and Representation

[NEXT] In terms of data collection and representation, one could argue that choosing which pieces of music to include in a corpus is a type of musical analysis, since it involves judgments on the timespan or styles under study. Because John Covach's paper later in the session will address this issue, I won't steal any of his thunder here, except to say that this is not an issue unique to pop/rock music.

What is more unique to a corpus study of pop/rock is that—unlike classical music—none of the musical content is explicitly given. [NEXT] In a classical work, such as the opening to this Courante by J. S. Bach, the score shows discrete pitch and rhythmic information. [NEXT] We can thus faithfully encode the work, such as in the music 21 format shown here. With popular music,

however, no official score usually exists aside from the original audio recording. Thus in contrast to Western art music, any symbolic representation of pop/rock music inherently requires an intermediary analytical stage. [NEXT] For example, here is my transcription of the melody for a passage from the song "I Can't Make You Love Me," as recorded by Bonnie Raitt. My transcription is a good approximation of the melody, but it is an approximation nonetheless. For comparison, here is a commercially-sold transcription of the same passage. [NEXT] The two versions are very similar, but some differences can be found, [NEXT] such as the tuning of the second note in the second measure and the timing of the last note in the second measure. In the second system, there is also a difference in the timing of the pentatonic descent from the D to the low F as well as the octave of the pitch F prior to the beginning of the chorus. Admittedly, these are small differences, but they show the difficulty of separating performance from composition with pop/rock music. My version leans more towards trying to capture the subtle tuning and timing aspects of Bonnie Raitt's performance, whereas the published transcription below leans more towards better reflecting the underlying composition of the melody. Here is what this sounds like [NEXT].

The subjective aspect of transcription is even greater with harmony, since harmonic analysis involves reducing a texture of many notes to a single chord symbol. For example, is the last chord in the first line a C minor 7 chord, as shown here, or does the D in the melody make it a C minor 9 chord? The answer depends on the disposition of the analyst. Even if we agree on the notes in the chord, we may disagree on its label. For example, there are at least two ways to analyze the B-flat over F chord in the second system. [NEXT] One way would be to call it a tonic chord in second inversion, namely a I6/4 chord. [NEXT] The second way would be to call it a cadential 6/4 chord, namely a V chord with a suspended 6th and 4th above the bass. Both are standard practice in the music theory, but each results in a different set of statistics. In the first reading, we find that a 2 chord, which ends the first line, moves to a 1 chord. In the second reading, we find that a 2 chord moves instead to a 5 chord. Any statistics on root motions will thus simply reflect back to us our own analytical predispositions.

This reflection of our own analytical framework back to us in our statistics is an important trap to avoid or acknowledge, but how? As a first step, it is critical to have more than one encoder involved in a corpus study of pop/rock music, and moreover, for each of those encoders to do their analyses independently. Doing so, we have the chance to assess subjectivity. [NEXT] For example, in my corpus with Davy Temperley of harmony in 100 rock songs, we agreed on the key or pitch center about 97% of the time. With regard to the absolute root of a chord, such as whether it was an A or a D chord, we agreed about 94% of the time. And in terms of the function or Roman numeral, we agreed about 92% of the time. The fact that our agreement was consistently above 90% seems good, but it's not 100%. Because there is currently no other corpus of pop/rock music that has multiple individual analyses, we don't know yet whether these figures are typical or atypical.

To be clear, I admit that some aspects of a pop/rock song can be objectively measured. [NEXT] For example, in their corpus study published in 2012, Schellenberg and von Scheve report statistics on Top 40 songs from the *Billboard* charts spanning 1965 to 2009. As this table shows, pop songs from the latter half of the 1980s were the longest, while songs from the late 1960s were the shortest.

We must be careful, though, not to overestimate how objectively any parameter can be measured [NEXT] For instance, Schellenberg and von Scheve also report that average tempo decreased from late 1965 to 2009. To the casual reader, this finding may seem unproblematic. After all, deciding whether a song is fast or slow may seem to some like a straightforward process. But recent perceptual studies of tempo in popular music show a great deal of variation between listeners. [NEXT] To illustrate this, I will use two versions of the song "Teardrops on My Guitar" by Taylor Swift. The first is a remix release called the "pop" version. Most listeners would presumably hear the tempo of this "pop" version as 100 BPM, with kick and snare corresponding to each beat in 4/4. [NEXT] [NEXT] Now let's listen to the original version. As you listen, note that the harmony and melody are going as fast as the first version but the drums are going at half speed. If a listener attends to the pacing of the harmonic and melodic content, they may feel the beat around 100 BPM. But if a listener attends to the pacing of the drums, they may feel the beat around 50 BPM—in other words, half as fast. Our notion that any given song has a single tempo, therefore, may itself be problematic. I'll go back and forth each measure conducting these two tempos. [NEXT]

To be fair, Schellenberg and von Scheve admit that measuring the tempo of a song was <quote> "complicated" (198). Their solution was to have two musicians each independently rate the tempo of every song, and if the tempo ratings did not agree, a third musician would resolve the disagreement. Any ambiguity in creating the corpus was thus completely removed from the final version. This is not the only corpus to use this method. The McGill *Billboard* corpus also used two musicians to independently analyze the chords for each song and then brought in a third "meta-annotator" to compare the two versions and decide what would be the final transcription. Here again, all information with regard to ambiguity was removed. The final corpus is thus presented as if its contents are entirely objective—that there is a single "best answer." Instead, I believe we should capture and investigate the extent to which and the situations in which analysts disagree. Untangling that, I would argue, is as important if not more important than any statistical results we obtain.

Embracing ambiguity is important because many researchers believe that a central benefit of corpus work on pop/rock music is to the lay the foundation for automated computer analysis of music. A human-annotated corpus is meant to act as the "ground truth" from which a computer will learn so as to be able to automatically analyze new music. This approach has practical uses for the music industry. [NEXT] For example, here is a screen shot from the "Sort Your Music" feature in Spotify. With this feature, users can sort their music on various parameters, such as release date, loudness, and BPM. Let's take a closer looks at the BPM parameter. [NEXT] Here are three songs and the BPM values that Spotify has identified. For the sake of time, I will play only the third song—"Out of Mind" by Colbie Caillat—which was rated at 180 BPM. As the song plays, I will clap the tempo that Spotify has identified. [NEXT]. As I hope you noticed, the algorithm does a good job synchronizing with the music, but there may be an "octave error," in that the tempo that probably most listeners would hear would be half as fast. In other words, the software is having trouble determining whether a song is fast or slow, which is a pretty basic aspect of how we traditionally think about tempo.

So work remains to be done, but how much? [NEXT] Research by Mark Levy offers some insight. The software driving Spotify's music analysis algorithm is the EchoNest API. In this table, Levy compares BPM ratings from the EchoNest API—shown in the top row—to crowd-sourced ratings

of the same songs. [NEXT] As you can see, the EchoNest API gets the same BPM value as the crowd-sourced value only about 40% of the time. [NEXT] In fact, about a quarter of the time, the EchoNest value is completely unrelated to the crowd-sourced value.

As it stands today, it seems, we should be wary of any corpus generated by a computer algorithm. For example, the "Million Song Dataset"—which I mentioned earlier—promises to be a great resource. Unfortunately, it was generated using the EchoNest API. So even though there are a million BPM ratings in the corpus, only about 40% correspond to a human listener's rating. The computer algorithm, in other words, reflects back not the perception of a human listener but rather the mechanics of its programming framework, much in the same way as the encodings of a human reflect back their own analytical framework.

Data Analysis and Interpretation

[NEXT] I will move to Part 2 concerning data analysis and interpretation. Because my time is limited here, I will avoid discussing the standard statistical fallacies we might find with any corpus study, such as confusing statistical significance with practical significance or correlation versus causation, since these types of statistical errors are well documented elsewhere. Instead, I address some pitfalls more specific to studies of pop/rock music, many of which derive from the underlying subjectivity in the encoding process.

[NEXT] For example, here is a table from Jay Summach's 2012 corpus study of the *Billboard* charts from 1965 to 1989. As this table shows, verse-chorus songs—both those without a prechorus (in the first row) and those with a prechorus (in the second row)—have on average been getting longer from 1965 to 1989. [NEXT] With that in mind, look at this next example also from Summach's study. Here, the white bars correspond to the total number of verse-chorus songs in the *Billboard* Top 20 per year. The green bars show the proportion of these verse-chorus songs that have prechorus sections. Note that—as shown by the percentage values in the bottom row—the distribution of verse-chorus songs either with or without prechorus sections changed over this period, such that most verse-chorus songs in the early 60s and 70s did not have prechorus sections whereas most verse-chorus songs in the late 1980s did have prechorus sections. Remember, though, that the average length of a song increased during this same period. If, therefore, an analyst has a threshold for how long a passage must be in order to be classified as a prechorus, we should not be surprised to find prechorus sections to be more common in longer songs. And in fact, Summach states in a 2011 article that a 4-bar passage is not long enough, in his opinion, to be a prechorus. The finding here that prechorus sections are more common in the late 1980s than in earlier years is thus arguably predicated on Summach's personal interpretation of what constitutes a prechorus.

Ultimately, the analysis of form is probably the most idiosyncratic element in music theory. We should be especially cautious, therefore, when we find statistics on form that derive from the analysis of only a single listener. [NEXT] Yet these sorts of statistics are fairly common in published articles, such as the data shown here from a corpus study by David Tough on recent *Billboard* songs. As Tough reports, 68% of his corpus—that's the blue and red portions of the pie chart together—have intro sections that last 10 seconds or less. For the sake of argument, let's assume that these figures reflect the hearing of all listeners. What are we to even make of this

finding? Tough posits that if you want a song to become commercially successful in today's market, the song should have a short intro section because that is how most modern hits are structured. This line of reasoning exemplifies a field known as "Hit Song Science." The presumption is that if a song is like other songs that are hits, then the song has a higher likelihood of becoming a hit itself. A number of data scientists have disproven the viability of this hypothesis, at least given the current state of research, but it's a philosophy that still underpins much work on popular music. The nagging question is whether a song achieves success because of its typicality or, conversely, because of its atypicality. Certainly, Nirvana's "Smells Like Teen Spirit" did not sound anything like the *Billboard* hits that preceded it, for example.

[NEXT] As I near the end of my talk, I want to reaffirm some of the windfalls of corpus work on pop/rock music, since I have highlighted many pitfalls. Traditionally speaking, corpus studies have been considered a subfield of music cognition. A corpus study is thus a way to understand human perception—to understand our built-in subjectivity and how that subjectivity affects our analyses. For example, here is data I presented at the last EuroMAC conference. This table shows the average duration of chords, in bars, for verse and chorus sections in the *Rolling Stone* magazine corpus I created with David Temperley. As you can see, both Temperley (DT) and I (TdC) have average chord lengths in verse sections that are longer than those in chorus sections. We might hypothesize, therefore, that a listener's perception of a section's role is affected by the durations of its chords. This finding appears to be relatively intersubjective, because it is reflected in both my analyses as well as Temperley's. We must be careful, though, with averages calculated across an entire body of songs because it may not represent any broad stylistic trait but rather the generic midpoint of multiple smaller populations. [NEXT] In fact, the reason that chord durations appear shorter in chorus sections than verse sections seems entirely predicated on the length of the tonic chord alone. There does not seem to be any significant difference in the length of non-tonic chords when comparing verse and chorus sections. Based on this data, I speculate that one factor that makes a passage sound more like a verse than a chorus derives from the length of the tonic harmony. My study of our corpus, therefore, has been a tool for me to interrogate my own perception, and in doing so, to compare it to the perception of others.

Conclusion

In conclusion, corpus work offers a number of windfalls for music research. Key-finding algorithms, for example, can be trained on symbolic data. Additionally, if the symbolic data is time-aligned with original recordings, it can provide a "ground truth" for computerized tempo estimation, chord extraction, and melodic transcriptions of raw audio files. We must be sure to recognize and reflect, however, the great variability in interpretation that exists between human analysts in these tasks, else we become trapped in pitfalls of our own making. In order to refine our corpus-based analytical methods, we must embrace this ambiguity, which mirrors our own perception and intrinsic preconceptions. Corpus work on pop/rock music can then and only then shed light not just on the music under study, but also perhaps more importantly on the analytical, theoretical, and perceptual frameworks that we use to encode this music.