Characterizing Musical Correlates of Large-Scale Discovery Behavior

Blair Kaneshiro [1] **Brandi Frisbie** [1] **Elena Georgieva** [1] **Daniel P. W. Ellis** [2]

Abstract

We seek to identify musical correlates of real-world discovery behavior by analyzing users' audio identification queries from the Shazam service. Recent research has shown that such queries are not uniformly distributed over the course of a song, but rather form clusters that may implicate musically salient events. We extend this research with the long-term goal of modeling query behavior computationally. Human annotators identified candidate musical events preceding increases in query volume. Categories derived from the annotations suggest that musical events relating primarily to structural segmentation boundaries motivate increased query likelihood.

1. Introduction

Audio content recognition services such as Shazam facilitate instantaneous music discovery in real-world encounters. Despite the widespread adoption of this behavior, however, it remains largely unstudied. A recent study by Kaneshiro et al. (2017) presented a first large-scale analysis of Shazam queries, with a focus on the 'offset'—the timing of queries within a song. They found that distributions of query offsets for 20 Billboard hit songs were not uniform over the course of a song, but rather exhibited prominent peaks with varying timings and shapes. Follow-up analyses showed that salient musical events common to all the songs—particularly the first entrance of vocals and first chorus—reliably drove increases in query volume. These findings suggest that further analysis of offset histograms in relation to corresponding audio content could provide valuable insights toward characterizing this ubiquitous musical behavior. In other words, what types of musical occurrences compel users to engage through discovery? As a step toward modeling this computationally, here we have conducted an annotation study to identify perceptually salient musical events preceding increases in query likelihood. We hypothesized that events related to structural segmentation, such as song-part boundaries, would play a role in motivating queries. Our findings, and insights from the annotation process itself, can inform future selection of audio features for computational modeling of query-based music discovery at scale.

[1]Stanford University, Stanford, USA [2]Google, Inc., New York, USA. Correspondence to: Blair Kaneshiro <blairbo@ccrma.stanford.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2. Methods

We analyzed the publicly available dataset (Shazam Entertainment, Ltd., 2017) released with Kaneshiro et al. (2017), containing over 188 million query offsets across the 20 songs. The relationship between music and queries was assumed to be causal—i.e., queries occurred in response to, not anticipation of, corresponding musical events, with an unknown latency in the range of seconds. Kaneshiro et al. (2017) assessed music-to-query relationships using percentile histogram slopes. Here we selected a slightly earlier feature—histogram minima—which can be thought to reflect points at which the global decrease in queries over time begins to reverse temporarily in response to some event. As in the previous study, we analyzed the first 1 million queries, by date, for each song. Query offset histograms were computed on a per-song basis in 1-second bins; local minima were identified after smoothing each histogram in 2-second bins. For added reference, song-part metadata obtained from Genius.com were manually aligned with the audio. Three trained musicians provided free-text annotations describing candidate musical events preceding each minimum, considering content between 2–10 seconds prior. Following annotation, four musical categories which exhaustively described the set of annotations were derived. The categories largely reflected varying levels of the structural segmentation hierarchy common to the pop genre. Song-part boundaries, as defined by the lyrics metadata, formed the top-level category; next were segmentation-based novelty events within song parts—i.e., notable changes at phrase boundaries, often related to instrumental texture or vocal register. The third category involved immediate, more exact repetitions of phrases, and the last comprised salient events not tied to segmentation boundaries, such as vocal riffs, standalone instrumental motifs, or utterances of song titles. As the categories were hierarchical, each minimum was coded to a single category by comparing it to each in order until the first appropriate match was identified.

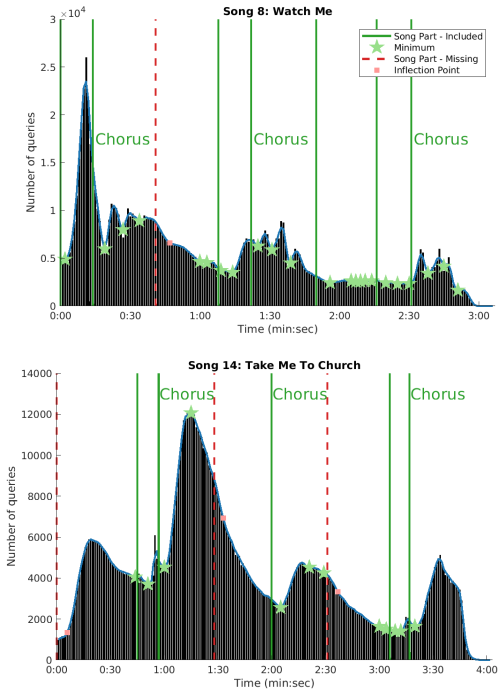

Figure 1. Example offset histograms.

3. Results and Discussion

A total of 404 histogram minima were detected and annotated across the 20 songs. As can be appreciated from Figure 1, minima preceded histogram peaks of varying heights and durations; and repeated occurrences of song parts at times produced similar patterns of peaks. Across all minima, song-part boundaries were coded most often (159 instances). However, not all song-part boundaries corresponded to histogram minima: The lyrics metadata identified an additional 28 boundaries, many of which could be related to non-minima inflection points in the histograms (Figure 1, red annotations). Of the remaining minima, 149 reflected phrase repetition; 49 reflected segmentation-based novelty events; and 47 reflected non-segmentation salience. When only the highest peak of each histogram was considered, song-part boundaries accounted for 12 songs, followed by repeated phrases (5 songs) and segmentation-based novelty (3 songs); there were no instances of non-segmentation salience here.

Computational modeling of human perception is challenging even when perception is unambiguous. In previous analyses of Shazam offsets, fundamental stimulus events were pre-defined and then assessed in the response. Our present annotation process, which worked backward from response to stimulus, highlighted perceptual ambiguities which could inform future computational endeavors. First, interpretations were often not unanimous. For example, when structural segmentation boundaries included both salient content concluding the current section and novel content introducing the subsequent section, coders sometimes disagreed as to which was the relevant musical event; this occurred at the level of both song-part boundaries and segmentation-based novelty events. Challenges also arose when annotating minima that were shallow or occurred in quick succession (as shown toward the endings of songs in Figure 1). These ambiguities and challenges were compounded by the fact that exact response latencies were not known, and were also found to vary across, and perhaps within, songs.

Our findings suggest a number of actionable insights. First, while annotations accounted for all histogram minima, the minima themselves accounted for only 85% of song-part boundaries. Therefore, it is highly likely that our approach failed to account for every occurrence of events from the other three categories as well. Future annotation endeavors could therefore use the now-established categories as a starting point, coding for *all* of their occurrences in the songs and subsequently reporting percentages, rather than counts, of event types preceding histogram minima. Such an adjustment could also clarify the large number of minima explained by phrase repetition—an event that will by definition occur more frequently than those implicating higher levels of the structural segmentation hierarchy. Alternative histogram features could take other inflection points into account or de-emphasize shallow dips, and histograms could be 'whitened' to control for the global decline in queries over time. While our present analysis assigned equal importance to all coded events, one could also weight each event by a prominence measure computed from corresponding histogram features (e.g., peak height or duration) in order to clarify their importance.

4. Conclusion

Expert annotation of large-scale music discovery data in conjunction with audio has revealed relevant categories of musical events driving increases in Shazam queries. These categories highlight the importance of structural segmentation boundaries at multiple time scales, as well as non-segmentation salient events. We additionally identified challenges and areas of perceptual ambiguity. As future work we will extend this analysis to computationally extracted audio features. Based on our findings, it appears that a combination of features capturing both novelty (e.g., Foote (2000)) and instantaneous salience (such as Vocal Prominence, Van Balen et al. (2015)) may model query behavior well. A direct comparison of time-varying audio and response features will also facilitate modeling of response latency. We will consider alternative response features and relate them to different levels of the structural segmentation hierarchy. Finally, we aim to characterize additional histogram attributes, such as morphology of the peaks.

Acknowledgments

We thank Alan Huang for coding the song-part metadata, and Nick Gang for helpful discussions about this work.

References

Foote, J. Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo*, volume 1, pp. 452–455, 2000.

Kaneshiro, B., Ruan, F., Baker, C. W., and Berger, J. Characterizing listener engagement with popular songs using large-scale music discovery data. *Frontiers in Psychology*, 8:416, 2017.

Shazam Entertainment, Ltd. Shazam Research Dataset—Offsets (SRD-O). In *Stanford Digital Repository*, 2017. URL https://purl.stanford.edu/fj396zz8014.

Van Balen, J. M. H., Burgoyne, J. A., Bountouridis, D., Müllensiefen, D., and Veltkamp, R. C. Corpus analysis tools for computational hook discovery. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pp. 227–233, 2015.