UTILIZING HIERARCHICAL STRUCTURE FOR AUDIO-BASED MUSIC SIMILARITY

Christos Plachouras
New York University Abu Dhabi
cplachouras@nyu.edu

ABSTRACT

In this work, I introduce a methodology for measuring the similarity between musical pieces by computing a hierarchical representation of their structure from their audio and comparing audio sections that have a similar structural function. Between a pair of musical pieces, the methodology aims to maximize how much of their audio is used to compute their similarity, under the constraint of only comparing structural segments that are deemed related. This introduces musical structure as a relevant characteristic for music similarity metrics, while minimizing the loss of information about the temporal evolution of music features within pieces. Experiments in music similarity measurements within musical genres as well as between studio and live performances are presented.

1. INTRODUCTION

1.1 Overview

One of the most active topics in the field of Music Information Retrieval is audio-based measurement of similarity between different musical pieces. This topic underlies many of the technologies we use daily, such as music recommendation systems relying on content-based approaches, cover song detection, genre classification, audio thumbnails, summaries, and fingerprints, and others.

Defining a musically-meaningful similarity metric between a sequence of audio features among pieces is not a trivial task. While some musical characteristics remain relatively constant throughout a musical piece, others vary significantly over time, and that evolution may be perceptually relevant.

The change of musical characteristics in a piece is often particularly evident during structural segment changes, where we can potentially observe larger, immediate changes in instrumentation, loudness, harmonic information, or others. If we wanted to compare a 10-second audio chunk from two rock pieces, our musical intuition might point to collecting that chunk from within a specific section of the piece, such as the chorus. If we selected that chunk randomly, we might end up with a chunk containing both part of the verse and part of the chorus. This means that we would be comparing a, potentially, comparatively invariant sequence of music features to one with at least one point of comparatively large change of music features. For brevity, audio chunks where at least one chunk contains more than one structural segment change will be referred to as 'non-aligned' for the rest of the paper.

1.2 Related work

Given the sequences of time-related features of musical pieces, a common first step for measuring their similarity is either truncating or padding them to the same length, or selecting a fixed-length segment from the start, middle, or end [1] [2]. Truncating results in loss of information to be compared and in, potentially, non-aligned chunks. Similarly, padding guarantees non-aligned chunks because of the introduction of 'silence' segments of different lengths. Audio thumbnailing has also been used [3] so that the selection is more representative of the piece, but large sections of audio are still ignored, and the thumbnails can still be non-aligned.

Another approach is using high-level, aggregate descriptors that characterize some features of the audio within a given dimension [4] or, more simply, using min, max, mean, or standard deviation of features [5]. The drawback of these methods is the loss of information about the temporal evolution of those features. Gaussian Mixture Models used for measuring spectral similarity similarly fail in this regard due to the random frame selection. Lastly, Hidden Markov Models [6] have shown mixed results, while Recurrent Neural Networks show more promise [7] [8].

2. METHODOLOGY

There are two main steps to this methodology. First, the structure of each piece needs to be analyzed at different levels of granularity. Secondly, rules for automatically selecting relevant segments to compare need to be defined according to our knowledge of the data.

2.1 Structural Analysis

Most music datasets do not contain human annotations of the structure of each piece. Additionally, evaluating

whether a structural segment is similar in function to a segment of another piece is not a clearly defined task. In fact, we might be interested in similarities of structure at different levels of granularity when perceiving music similarity, from the level of repeating sequences and lyric lines, to the overarching structure of the dynamics and timbre.

As a solution to these problems, the hierarchical structure decomposition method by McFee and Ellis [9] was used. This method uses a similarity graph to encode global repetition and local consistency in a sequence of time-related music features. Spectral clustering using a range of component values, which correspond to the number of desired segment types, segments the graph at different levels of granularity. The resulting hierarchical structure of two songs is shown in Fig 1.

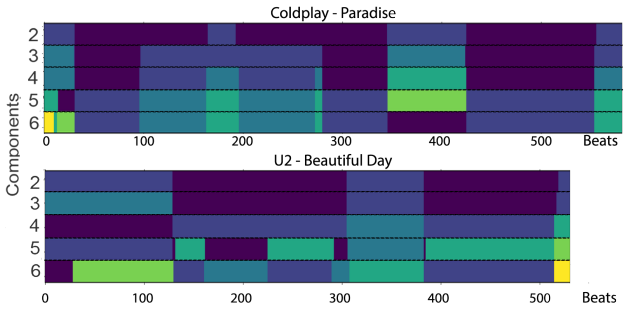

Figure 1. Hierarchical structure with $m = 2, 3, 4, 5, 6$ components

2.2 Segment selection

In its simplest form, the selection process starts with beat-synchronizing the feature sequences. Then, across the structural hierarchies of two musical pieces, we pair up all structural segments that have the same number of beats. Since all beat-synchronized paired segments will have the same length, we can use a distance metric to measure the total dissimilarity between all pairs.

This initial selection process is not applicable to every music style, but it is more flexible than it seems. Since we are checking across all levels of structure granularity, we can be pairing, for example, the verse of one piece to an equal-length part of the verse of another. Importantly, we are favoring comparisons between segments that are comparatively invariant in music characteristics and minimizing comparisons between non-aligned audio chunks.

There are further improvements to the core selection process depending on our knowledge of the data to improve results and reduce the computational cost. A segment is most likely going to be paired with multiple other segments. We can instead choose to only keep the closest-positioned pair. For example, if a segment from song A that starts at the middle of the song is paired with a segment from song B starting at the start and one segment starting slightly after the middle of song B, we would only keep the latter pair. Another simple rule would be to only compare the largest segments that were paired, in case we know that small segments will be insignificant in the comparison.

3. EXPERIMENTS

To understand the behavior of this methodology, we will look into its application in different musical contexts that are constituent to larger tasks involving music similarity measurement. The source code, including the implementation details, for these experiments is publicly available.

We will first look into two very popular songs loosely within the genre of pop rock: *Paradise* by Coldplay and *Beautiful Day* by U2. We can see the hierarchical structure in Fig. 1

After selecting pairs of segments with the same number of beats across the structural hierarchies, we find 32 segment pairs. in Fig. 2, for visualization purposes, we merge overlapping segments to more clearly distinguish that most of each song was selected for comparison.

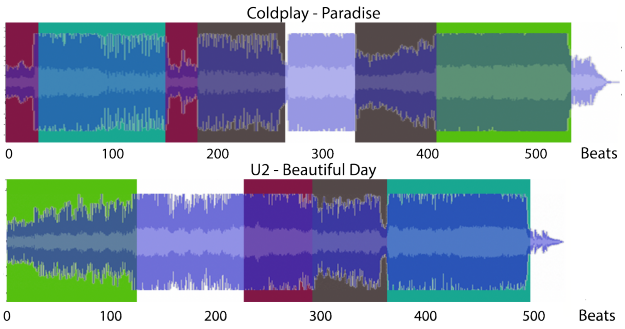

Figure 2. Matched structural segments (pairs indicated with the same color)

We apply the same methodology to compare the studio version of *Paradise* to its live performances in Paris, Glastonbury, and a Late Show. Due to space constraints, Fig. 3 is further simplified to only show what parts of the songs are selected. Again, in all three comparisons, the majority of the audio was selected for comparison.

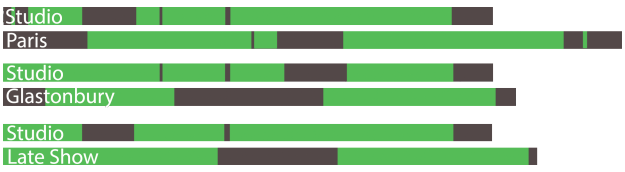

Figure 3. Matched structural segments (valid sections in green)

4. FUTURE WORK

These experiments show promising results, as less information is sacrificed and comparison between non-aligned audio chunks is avoided. Further experimentation and quantitative evaluation of it as well as of the segment selection rules in different tasks and datasets related to music similarity are needed to understand its applicability, behavior, and performance.

5. REFERENCES

[1] K. Trochidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music into emotions," vol. 2011, 01 2008, pp. 325–330.

[2] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "A machine learning approach to automatic music genre classification," *Journal of the Brazilian Computer Society*, vol. 14, pp. 7 – 18, 09 2008. [Online]. Available: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-65002008000300002&nrm=iso

[3] D. F. Silva, C.-C. M. Yeh, G. E. A. P. A. Batista, and E. J. Keogh, "Simple: Assessing music similarity using subsequences joins," in *ISMIR*, 2016.

[4] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011. [Online]. Available: https://doi.org/10.1121/1.3642604

[5] M. Slaney, K. Q. Weinberger, and W. White, "Learning a metric for music similarity," in *ISMIR*, 2008.

[6] A. Flexer, E. Pampalk, and G. Widmer, "Hidden markov models for spectral similarity of songs," 2005.

[7] A. Balakrishnan, "Deepplaylist : Using recurrent neural networks to predict song similarity," 2016.

[8] M. Jiang, Z. Yang, and C. Zhao, "What to play next? a rnn-based music recommendation system," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, 2017, pp. 356–358.

[9] B. McFee and D. P. W. Ellis, "Analyzing song structure with spectral clustering," in *ISMIR*, 2014.