COGNITION-INSPIRED DESCRIPTORS FOR SCALABLE COVER SONG RETRIEVAL

Jan Van Balen, Dimitrios Bountouridis, Frans Wiering, Remco Veltkamp
Utrecht University, Department of Information and Computing Sciences
j.m.h.vanbalen@uu.nl, d.bountouridis@uu.nl

ABSTRACT

Inspired by representations used in music cognition studies and computational musicology, we propose three simple and interpretable descriptors for use in mid- to high-level computational analysis of musical audio and applications in content-based retrieval. We also argue that the task of scalable cover song retrieval is very suitable for the development of descriptors that effectively capture musical structures at the song level. The performance of the proposed descriptions in a cover song problem is presented. We further demonstrate that, due to the musically-informed nature of the descriptors, an independently established model of stability and variation in covers songs can be integrated to improve performance.

1. INTRODUCTION

The aim of this paper is to demonstrate the use of three new, cognition-inspired music descriptors for content-based music retrieval.

1.1 Audio Descriptors

There is a growing consensus that some of the most widely used features in Music Information Research, while very effective for engineering applications, do not serve the dialog with other branches of music research [1]. As a classic example, MFCC features can be shown to predict human ratings of various perceptual qualities of a sound. Yet, from the perspective of neuropsychology, claims that they mathematically approximate parts of auditory perception have become difficult to justify as more parts of the auditory pathway are understood.

Meanwhile, a recent analysis of evaluation practices by Sturm [18] suggests that MIR systems designed to classify songs into high-level attributes like genre, mood or instrumentation may rely on confounded factors unrelated to any high-level property of the music, even if their performance numbers approach 100%. Researchers have focused too much on the same evaluation measures and the same datasets and as a result, today, top performing genre and mood recognition systems rely on the same low-level features that are used to classify bird sounds. [1]

We also observe that, despite the increasing availability of truly big audio data and the promising achievements of MIR over the last decade, studies that turn big audio data into findings about music itself seem hard to find. Notable exceptions include studies on scales and intonation, and [16]. In the latter, pitch, timbre and loudness data were analyzed for the Million Song Dataset, focusing on the distribution and transitions of discretized code words. Yet, we have also observed that this analysis sparks debate among music researchers outside the MIR field, in part because of the descriptors used. The study uses the Echo Nest audio features provided with the dataset, which are computed using undisclosed, proprietary methods and therefore objectively difficult in interpretation.

1.2 Towards Cognitive Audio Descriptors

In a longer-term effort towards modeling cognition level qualities of music such as its complexity, expectedness and repetitiveness from raw audio data, we aim to design and evaluate features that describe harmony, melody and rhythm on a level that has not gained the attention it deserves in MIR's audio community, perhaps due to the 'success' of low-level features discussed above. In the long run, we believe, this will provide insights into the building blocks of music: riffs, motives, choruses, and so on.

1.3 Cover Song Detection

In this section, we argue that the task of scalable cover song retrieval is very suitable for developing descriptors that effectively capture mid- to high-level musical structures, such as chords, riffs and hooks.

Cover detection systems take query song and a database and aim to find other versions of the query song. Since many real-world cover versions drastically modulate multiple aspects of the original: systems must allow for deviations in key, tempo, structure, lyrics, harmonisation and phrasing, to name just a few. Most successful cover detection algorithms are built around a two-stage architecture. In the first stage, the system computes a time series representation of the harmony or pitch for each of the songs in a database. In the second stage, the time series representing

[1] largely MFCC and spectral moments, see [6, 18] for examples

the query is compared to each of these representations, typically by means of some kind of alignment, i.e. computing the locations of maximum local correspondence between the two documents being compared. See [15] for more on this task and an overview of cover detection strategies.

2. SCALABLE COVER SONG RETRIEVAL

Generally, alignment methods are computationally expensive but effective. Results achieved this way have reached mean average precision (MAP) figures of around 0.75 at the MIREX evaluation exchange. [2]

When it comes to large-scale cover detection (hundreds of queries and thousands of songs), however, alignment-based methods can become impractical. Imagine a musicologist whose aim is not to retrieve matches to a single query, but to study all the relations in a large, representative corpus. Alignment-based techniques are no longer an option: a full pair-wise comparison of $10,000$ documents would take weeks, if not months. [3] .

This is why some researchers have recently focused on scalable techniques for cover song detection. Scalable strategies are often inspired by audio fingerprinting and involve the computation of an indexable digest of (a set of) potentially stable landmarks in the time series, which can be stored and matched through a single inexpensive look-up. Examples include the 'jumpcodes' approach by [2], the first system to be tested using the Million Song Dataset. This study reports a recall of 9.6% on the top 1 percent of retrieved candidates. Another relevant example is the interval-gram approach by Walters [19], which computes fingerprinting-inspired histograms of local pitch intervals, designed for hashing using wavelet decomposition.

Reality shows that stable landmarks are relatively easy to find when looking for exact matches (as in fingerprinting), but hard to find in real-world cover songs. A more promising approach was presented by Bertin-Mahieux in [3]], where the 2D Fourier transform of beat-synchronized chroma features is used as the primary representation. The accuracy reported is several times better than for the system based on jumpcodes. Unfortunately, what exactly these features capture is difficult to explain.

The challenges laid out in the above paragraph make cover song detection an ideal test case to evaluate a special class of descriptors: harmony, melody and rhythm descriptors, global or local, which have a fixed dimensionality and some tolerance to deviations in key, tempo and global structure. If a collection of descriptors can be designed that accurately describes a song's melody, harmony and rhythm in a way that is robust to the song's precise structure, tempo and key, we should have a way to determine similarity between the 'musical material' of two songs and assess if the underlying composition is likely to be the same.

[2] `http://www.music-ir.org/mirex/wiki/2009:Audio_Cover_Song_Identification_Results`

[3] MIRex 2008 (the last to report runtimes) saw times of around $1.4 - 3.7 \times 10^5$ s for a task that involves $115,000$ comparisons. The fastest of these algorithms would take 1.8 years to compute the $\frac{1}{2}10^8$ comparisons required in the above scenario. The best performing algorithm would take 6 years.

3. PITCH AND HARMONY DESCRIPTORS

There is an increasing amount of evidence that the primary mechanism governing musical expectations is statistical learning [7, 12]. On a general level, this implies that the conditional probabilities of musical events play a large role in their cognitive processing. Regarding features and descripors, it justifies opportunities of analyzing songs and corpora in terms of probabily distributions. Expectations resulting from the exposure to statistical patterns have in turn been shown to affect the perception of melodic complexity and familiarity. See [7] for more on the role of expectation in preference, familiarity and recall.

We propose three new descriptors: the pitch bihistogram, the chroma correlation coefficients and the harmonization feature. The pitch bihistogram describes melody and approximates a histogram of pitch bigrams. The chroma correlation coefficients relate to harmony. They approximate the co-occurrence of chord notes in a song. The third representation, the harmonization feature, combines harmony and melody information. These three descriptors will now be presented in more detail.

3.1 The Pitch Bihistogram

Pitch bigrams are ordered pairs of pitches, similar to word or letter bigrams used in computational linguistics. Several authors have proposed music descriptions based on pitch bigrams, most of them from the domain of cognitive science [10, 11, 13]. Distributions of bigrams effectively encode first-degree expectations. More precisely: if the distribution of bigrams in a piece is conditioned on the first pitch in the bigram, we obtain the conditional frequency of a pitch given the one preceding it.

The first new feature we introduce will follow the bigram paradigm. Essentially, it captures how often two pitches p_1 and p_2 occur less than a distance d apart.

Assume that a melody time series $P(t)$, quantized to semitones and folded to one octave, can be obtained. If a pitch histogram is defined as:

$$h(p) = \sum_{P(t)=p} \frac{1}{n}, \tag{1}$$

with n the length of the time series and $p \in \{1, 2, \dots 12\}$, the proposed feature is then defined:

$$B(p_1, p_2) = \sum_{\substack{P(t_1)=p_1 \\ P(t_2)=p_2}} w(t_2 - t_1) \tag{2}$$

where

$$w(x) = \begin{cases} \frac{1}{d}, & \text{if } 0 < x < d. \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This will be reffered to as the **pitch bihistogram**, a bigram representation that can be computed from continuous melodic pitch. Note that the use of pitch classes rather than pitch creates an inherent robustness to octave errors in the melody estimation step, making the feature insensitive to one of the most common errors encountered in pitch extraction.

Alternatively, scale degrees can be used instead of absolute pitch class. In this scenario, the pitch contour $P(t)$ must first be aligned to an estimate of the piece's overall tonal center. As a tonal center, the tonic can be used. However, for extra robustness to misestimating the tonic, we suggest to use the tonic for major keys and the minor third for minor keys.

3.2 Chroma Correlation Coefficients

The second feature representation we propose focuses on vertical rather than horizontal pitch relation. It encodes which pitches appear simultaneously in a signal.

$$C(p_1, p_2) = \text{corr}(c(t, p_1), c(t, p_2)), \tag{4}$$

where $c(t, p)$ is a 12-dimensional chroma time series (also known as pitch class profile) computed from the song audio. From this chroma representation of the song $c(t, p)$ we compute the correlation coefficients between each pair of chroma dimensions to obtain a 12×12 matrix of **chroma correlation coefficients** $C(p_1, p_2)$. Like the pitch bihistogram, the chroma features can be transposed to the same tonal center (tonic or third) based on an estimate of the overall or local key.

3.3 Harmonisation Feature

Finally, the **harmonisation feature** is a set of histograms of the harmonic pitches $p_h \in \{1, \ldots, 12\}$ as they accompany each melodic pitch $p_m \in \{1, \ldots, 12\}$. It is computed from the pitch contour $P(t)$ and a chroma time series $c(t, p_h)$, which should be adjusted to have the same sampling rate and aligned to a common tonal center.

$$H(p_m, p_h) = \sum_{P(t)=p_m} c(t, p_h). \tag{5}$$

From a memory and statistical learning perspective, the chroma correlation coeffiencients and harmonisation feature may be used to approximate expectations that include: the expected consonant pitches given a chord note, the expected harmony given a melodic pitch, and the expected melodic pitch given a chord note. Apart from [8], where a feature resembling the chroma correlation coefficients is proposed, information of this kind has yet to be exploited in a functioning (audio) MIR system. Like the pitch bihistogram and the chroma correlation coefficients, the harmonisation feature has a dimensionality of 12×12.

4. EXPERIMENTS

To evaluate the performance of the above features for cover song retrieval, we set up a number of experiments around the *covers80* dataset by Ellis [5]. This dataset is a collection of 80 cover song pairs, divided into a fixed list of 80 queries and 80 candidates. Though *covers80* is not actually 'large-scale', it is often used for benchmarking [4] and its associated audio data are freely available. In contrast, the much larger Second Hand Songs dataset is distributed only in the form of standard Echo Nest features. These features do not include any melody description, which is the basis for the descriptors proposed in this study.

Regarding scalability, we chose to follow the approach taken in [19], in which the scalability of the algorithm follows from the simplicity of the matching step. The proposed procedure is computationally scalable in the sense that, with the appropriate hashing strategy, matching can be performed in constant time with respect to the size of the database. Nevertheless, we acknowledge that the distinguishing power of the algorithm must be assessed in the context of much more data. A large scale evaluation of our algorithm, adapted to an appopriate dataset and extended to include hashing solutions and indexing, is planned as future work.

4.1 Experiment 1: Global Fingerprints

In the first experiment, the three descriptors from section 3 were extracted for all 160 complete songs. Pitch contours were computed using Melodia and chroma features using HPCP, using default settings [14]. [5] For efficiency in computing the pitch bihistogram, the pitch contour was median-filtered and downsampled to $\frac{1}{4}$ of the default frame rate. The bihistogram was also slightly compressed by taking its square root.

The resulting reprentations (B, C and H) were then scaled to the same range by whitening them for each song individually (subtracting the mean of their n dimensions, and dividing by the standard deviation; $n = 144$). To avoid relying on key estimation, features in this experiment were not aligned to any tonal center, but transposed to all 12 possible keys. In a last step of the extraction stage, the features were scaled with a set of dedicated weights $w = (w_1, w_2, w_3)$ and concatenated to 12 432-dimensional vectors, one for each key. We refer to these vectors as the *global fingerprints*.

In the matching stage of the experiment, the distances between all queries and candidates were computed using a cosine distance. For each query, all candidates were ranked by distance. Two evaluation metrics were computed: *recall at 1* (the proportion of covers retrieved among the top 1 result for each query; R_1) and *recall at 5* (proportion of cover retrieved 'top 5'; R_5).

4.2 Experiment 2: Thumbnail Fingerprints

In a second experiment, the songs in the database were first segmented into structural sections using structure features as described by Serra [17]. This algorithm performed best at the 2012 MIREX evaluation exchange in the task of 'music structure segmentation', both for boundary recovery and for frame pair clustering. (A slight simplification was made in the stage where sections are compared: no dynamic time warping was applied in our model.) From this segmentation, two non-overlapping thumbnails are selected as follows:

[4] results for this dataset have been reported by at least four authors [15]

[5] mtg.upf.edu/technologies

1. Simplify the sequence of section labels (e.g. ababCabCC): merge groups of section labels that consistently appear together (e.g. AACACC for the example above).

2. Compute the total number of seconds covered by each of the labels A, B, C... and find the two section labels covering most of the song.

3. Return the boundaries of the first appearance of the selected labels.

The fingerprint as described above was computed for the full song as well as for the resulting thumbnails, yielding three different fingerprints: one global and two *thumbnail fingerprints*, stored separately. As in experiment 1, we transposed these thumbnails to all keys, resulting in a total of 36 fingerprints extracted per song: 12 for the full song, 12 for the first thumbnail and 12 for the second thumbnail.

4.3 Experiment 3: Stability Model

In the last experiment, we introduced a model of stability in cover song melodies. This model was derived independently, through analysis of a dataset of annotated melodies of cover songs variations. Given the melody contour for a song section, the model estimates the stability at each point in the melody. Here, stability is defined as the probability of the same pitch appearing in the same place in a performed variation of that melody.

The stability estimates produced by the model are based on three components that are found to correlate with stability: the duration of notes, the position of a note inside a section, and the pitch interval. The details of the model and its implementation are described in the following section.

5. STABILITY MODEL

The model we apply is a quantitative model of melody stability in cover songs. As it has been established for applications broader than the current study, it is based on a unique, manually assembled collection of annotated cover songs melodies. The dataset contains four transcribed melodic variations for 45 so-called 'cliques' of cover songs, a subset of the Second Hand Songs dataset. [6] . Some songs have one section transcribed, some have more, resulting in a total of 240 transcriptions.

For the case study presented here, the transcriptions were analysed using multiple sequence alignment (MSA) and a probabilistic definition of stability.

5.1 Multiple Sequence Alignment

MSA is a bioinformatics method that extends pairwise alignment of symbolic arrays to a higher number of sequences [4]. There are many approaches to MSA, some employing hidden markov models or genetic algorithms. The most popular approach is progressive alignment. This technique creates an MSA by combining pairwise alignments (PWA)

[6] `http://labrosa.ee.columbia.edu/millionsong/secondhand`

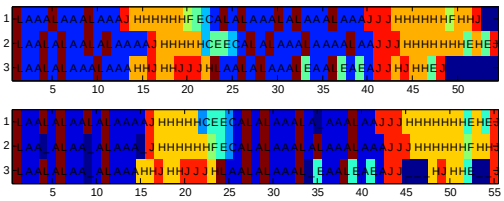

Figure 1. A clique of melodies before (top) and after (bottom) multiple sequence alignment.

starting from the most similar sequences, constructing a tree usually denoted as the 'guide tree'. Unlike MSA, pairwise alignment has been researched extensively in the (symbolic) MIR community, see [9] for an overview.

Whenever two sequences are aligned, a *consensus* can be computed, which can be used for the alignment connecting the two sequences to the rest of the three. The consensus is a new compromise sequence formed using heuristics to resolve the ambiguity at non-matching elements. These heuristics govern how gaps propagate through the tree, or whether 'leaf' or 'branch' elements are favored. The current model favors gaps and branch elements.

When the root consensus of the tree is reached, a last iteration of PWA's aligns each sequence to the root consensus to obtain the final MSA. Figure 1 shows two sets of melodic sequences (mapped to a one-octave alphabet $\{\mathrm{A} \ldots \mathrm{L}\}$) before and after MSA. Note that the MSA is based on a PWA strategy which maximizes an optimality criterion based on not just pitch but also duration and onset times.

5.2 Stability

The **stability** of a note in a melody is now defined as the probability of the same note being found in the same position in an optimally aligned variation of that melody.

Empirically, given a set of N aligned sequences

$$\{s_k(i)\} \quad i = 1 \ldots n,\ k = 1 \ldots N \tag{6}$$

we compute the stability of event $s_k(i)$ as:

$$stab(s_k(i)) = \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq k}}^{j=N} s_j(i) == s_k(i) \tag{7}$$

As an example, in a position i with events $s_1(i) = \mathrm{A}$, $s_2(i) = \mathrm{A}$, $s_3(i) = \mathrm{A}$ and $s_4(i) = \mathrm{B}$, the stability of A is 0.66. The stability of B is 0.

5.3 Findings

As described in the previous section, we drew a random sample of notes from the dataset in order to observe how stability behaves as a function of the event's pitch, duration and position inside the song section.

The first relationship has 'position' as the independent variable and describes the stability as it evolves throughout the section. Figure 2 shows how stability changes with

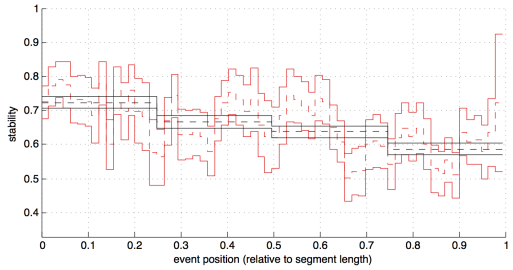

Figure 2. Stability of an event vs. position in the melody.

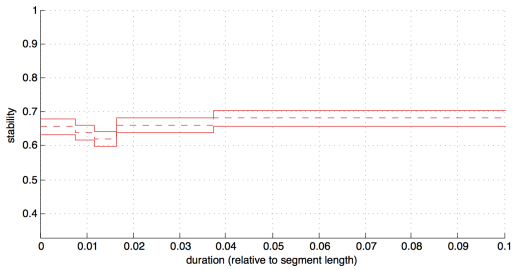

Figure 3. Stability of an event vs. duration.

position. The mean and 95% CI for the mean are shown for two different binnings of the position variable. The 4-bin curve illustrates how stability generally decreases with position. The more detailed 64-bin curve shows how the first two thirds of a melody are more stable than the last, though an increased stability can be seen at the end of the section.

Figure 3 shows the stability of notes as a function of their duration. The distribution of note durations is centered around 1% of the segment length. Below and above this value, the stability goes slightly up. This suggests that notes with less common durations are more stable. However, the trend is weak compared with the effect of position. Note duration information will therefore not be used in the experiments in this study.

Figure 4 shows the stability (mean and 95% CI for the mean) of a note given the pitch interval that follows. Note how the relative stability of one-semitone jumps stands out compared to repetitions and two-semitone jumps, even though two-semitone jumps are far more frequent. This suggests again that less-frequent events are more stable. More analysis as to this hypothesis will be performed in a later study.

6. DISCUSSION

Table 1 summarizes the results of the cover song retrieval experiments.

In the experiments where each descriptor was tested individually, the harmony descriptors (chroma correlation coefficients) performed best: we obtained an accuracy of over 30%. When looking at the top 5, there was a recall of 53.8%. The *recall at 5* evaluation measure is included to give an impression of the performance that could

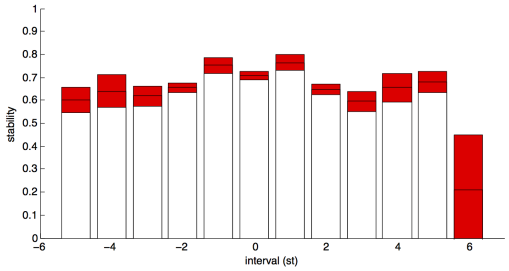

Figure 4. Stability of an event vs. the interval that follows.

be gained if the current system were complemented with an alignment-based approach to sort the top-ranking candidates, as proposed by [19].

The next results show that, for the three features together, the global fingerprints outperform the thumbnail fingerprints (42.5% vs. 37.5%), and combining both types does not increase performance further. In other configurations, thumbnail fingerprints were observed to outperform the global fingerprints. This is possibly the result of segmentation choices: short segments produce sparse fingerprints, which are in turn farther apart in the feature space than 'dense' fingerprints.

In experiment 3, two components of the stability model were integrated in the cover detection system. The 4-bin stability vs. position curve (scaled to the $[0, 1]$ range) was used as a weighting to emphasize parts of the melody before computing the thumbnails' pitch bihistogram. The stability per interval (compressed by taking its square root) was used to weigh the pitch bihistogram directly.

With the stability information added to the model, the top 1 precision reaches 45.0%. The top 5 recall is 56.3%. This result is situated between the accuracy of the first alignment-based strategies (42.5%), and the accuracy of a recent scalable system (53.8%; [19]). We conclude that the descriptors capture enough information to discriminate between individual compositions, which we set out to show.

7. CONCLUSIONS

In this study, three new audio descriptors are presented. Their interpretation is discussed, and results are presented for an application in cover song retrieval. To illustrate the benefit of feature interpretability, an independent model of cover song stability is integrated into the system.

We conclude that current performance figures, though not state-of-the-art, are a strong indication that scalable cover detection can indeed be achieved using interpretable, cognition-inspired features. Second, we observe that the pitch bihistogram feature, the chroma correlation coefficients and the harmonisation feature capture enough information to discriminate between individual compositions, proving that they are at the same time meaningful and highly informative, a scarse resource in the MIR feature toolkit. Finally, it has been demonstrated that the problems of cognition-level audio description and scalable cover detection can be succesfully addressed together.

	Descriptor	R_1	R_5
Global fingerprints	B	0.288	0.438
	C	0.313	0.538
	H	0.200	0.375
	$w = (2, 3, 1)$	0.425	0.575
Thumbnail fingerprints	$w = (2, 3, 1)$	0.388	0.513
Global + thumbnail fingerprints	$w = (2, 3, 1)$	0.425	0.538
Both fingerprints + stability model	$w = (2, 3, 1)$	0.450	0.563

Table 1. Summary of experiment results. w are the feature weights. Performance measures are *recall at 1* (proportion of covers retrieved 'top 1'; R_1) and *recall at 5* (proportion of cover retrieved among 'top 5'; R_5).

As future work, tests will be carried out to assess the discriminatory power of the features when applied to a larger cover song problem.

8. ACKNOWLEDGEMENTS

This research is supported by the NWO CATCH project COGITCH (640.005.004), and the FES project COMMIT/.

9. REFERENCES

[1] Jean-Julien Aucouturier and Emmanuel Bigand. Seven problems that keep MIR from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems*, 41(3):483–497, July 2013.

[2] T Bertin-Mahieux and Daniel P W Ellis. Large-scale cover song recognition using hashed chroma landmarks. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 10–13, 2011.

[3] T Bertin-Mahieux and Daniel P W Ellis. Large-Scale Cover Song Recognition Using The 2d Fourier Transform Magnitude. In *Proc Int Soc for Music Information Retrieval Conference*, pages 2–7, 2012.

[4] H Carrillo and D Lipman. The Multiple Sequence Alignment Problem in Biology. *SIAM Journal on Applied Mathematics*, 1988.

[5] Daniel P. W. Ellis and C.V. Cotton. The "covers80" cover song data set, 2007.

[6] M. Graciarena, M. Delplanche, E. Shriberg, A Stolcke, and L. Ferrer. Acoustic front-end optimization for bird species recognition. In *IEEE int conf on Acoustics Speech and Signal Processing (ICASSP)*, pages 293–296, March 2010.

[7] David Huron. Musical Expectation. In *The 1999 Ernest Bloch Lectures*. 1999.

[8] Samuel Kim and Shrikanth Narayanan. Dynamic chroma feature vectors with applications to cover song identification. *2008 IEEE 10th Workshop on Multimedia Signal Processing*, pages 984–987, October 2008.

[9] Peter van Kranenburg. *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*. PhD thesis, Utrecht University, 2010.

[10] Y. Li and D. Huron. Melodic modeling: A comparison of scale degree and interval. In *Proc. of the Int. Computer Music Congerence*, 2006.

[11] Daniel Müllensiefen and Klaus Frieler. Evaluating different approaches to measuring the similarity of melodies. *Data Science and Classification*, 2006.

[12] Marcus T. Pearce, Mara Herrojo Ruiz, Selina Kapasi, Geraint A. Wiggins, and Joydeep Bhattacharya. Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, 50(1):302 – 313, 2010.

[13] Pablo H Rodriguez Zivic, Favio Shifres, and Guillermo a Cecchi. Perceptual basis of evolving Western musical styles. *Proceedings of the National Academy of Sciences of the United States of America*, 110(24):10034–8, June 2013.

[14] J. Salamon and E. Gomez. Melody extraction from polyphonic music signals using pitch contour characteristics. In *IEEE Trans. on Audio, Speech and Language Processing*, 2010.

[15] Joan Serrà. *Identification of versions of the same musical composition by processing audio descriptions*. PhD thesis, Universitat Pompeu Fabra, 2011.

[16] Joan Serrà, Alvaro Corral, Marián Boguñá, Martín Haro, and Josep Ll Arcos. Measuring the evolution of contemporary western popular music. *Scientific reports*, 2:521, January 2012.

[17] Joan Serra, M Meinard, Peter Grosche, and Josep Ll Arcos. Unsupervised Detection of Music Boundaries by Time Series Structure Features. *Proc of the AAAI Conf on Artificial Intelligence*, pages 1613–1619, 2012.

[18] Bob L. Sturm. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, July 2013.

[19] Thomas C Walters, David A Ross, and Richard F Lyon. The Intervalgram : An Audio Feature for Large-scale Melody Recognition. In *Int. Symp. on Computer Music Modeling and Retrieval (CMMR)*, pages 19–22, 2012.