STRUCTURAL CHANGE ON MULTIPLE TIME SCALES AS A CORRELATE OF MUSICAL COMPLEXITY

Matthias Mauch **Mark Levy**
Last.fm, Karen House, 1–11 Bache's Street,
London, N1 6DL. United Kingdom.
matthias@last.fm mark@last.fm

ABSTRACT

We propose the novel audio feature *structural change* for the analysis and visualisation of recorded music, and argue that it is related to a particular notion of musical complexity. Structural change is a meta feature that can be calculated from an arbitrary frame-wise basis feature, with each element in the structural change feature vector representing the change of the basis feature at a different time scale. We describe an efficient implementation of the feature and discuss its properties based on three basis features pertaining to harmony, rhythm and timbre. We present a novel flower-like visualisation that allows us to illustrate the overall structural change characteristics of a piece of audio in a compact way. Several examples of real-world music and synthesised audio exemplify the characteristics of the structural change feature. We present the results of a web-based listening experiment with 197 participants to show the validity of the proposed feature.

Keywords: audio, musical complexity, visualisation

1. INTRODUCTION

A piece of music has many qualities that influence how it is perceived by human beings. These qualities include timbre, rhythm and harmony. One further, distinct property is the way in which timbre, rhythm, harmony and other features are temporally organised into units of various lengths over the course of the piece, from the smallest note change to the change between two sections. In this paper we propose an audio feature aimed at characterising part of this temporal, structural organisation.

A measure of structural change can be useful for music browsing within a track or in collections of music. In particular, suitable visualisations of the feature can directly be used for concise thumbnail-like descriptions of musical pieces. As a measure of complexity, structural change lends itself to the exploration of the cultural evolution of music.

Parry [8] provides an overview of research in music complexity and applies several measures of complexity on symbolic music. In the audio domain, Streich [10] gives a comprehensive description of existing theories and techniques. He also discusses many definitions of complexity in science and their application to music, noting that pure information-theoretical and mathematical approaches such as entropy and Kolmogorov complexity can limit the exploration of human-perceived complexity.

Our approach is inspired by a biological notion of complexity [1] according to which things are defined as more complex the less likely they could have come into existence by chance. More specifically, we focus on the aspect of distinction, the fact that "different parts of the complex behave differently" [5]. As an example in the domain of audio, consider two ten-second waveforms: one exclusively consisting of pink noise, the other one consisting of five seconds of pink noise followed by five seconds of white noise. Clearly, something must have happened in the middle of the second waveform that resulted in this change, or, in musical terms, the second piece must have had a 'composer'.

In real music, such structural changes happen in all musical qualities (including rhythm and harmony), and—equally importantly—they happen on all time scales within the range of the length of a piece. Our proposed feature captures these structural changes at several time scales. Our assumption is that it correlates with the degree to which the music was composed, an indication of complexity.

We would like to stress that the structural change feature is unrelated to any *instantaneous* complexity listeners may perceive. The timbre of a complete orchestra playing the same note, or the harmony of a rare jazz chord may sound complex, but our method exclusively aims at discovering the quantity of change.

Given an arbitrary audio feature (for example chroma), calculated for short frames across a piece of music, our proposed method calculates a meta-feature at every frame by

comparing statistics of the feature in a window before the current frame with statistics of a window after the current frame, i.e. a it compares left to right. This method resembles Foote's convolution with a checkerboard kernel [2], which is used for structural segmentation. Our approach focuses on the amount of change *itself* as a valid property of music. It is more similar in scope to Streich's tonal complexity measure [10, Chapter 4], which compares the harmonic content in one short-term window to that in a longer window. However, we are concerned with multiple time scales, and in order to capture the structural changes at different time scales this calculation is done for several different window sizes, resulting in a vector-valued feature.

There has been previous research in multi-time-scale analysis of audio properties, most prominently the keyscapes proposed by Sapp [9] and extensions thereof [4]. These analyses are aimed at providing information about what classes of harmonies are present in the signal at different time scales. While a visualisation of these classes may reveal changes in the signal, our proposed feature is concerned with the *amount* of change in any kind of frame-wise audio feature. In short, our approach combines Foote's measure of change with Sapp's multi-time-scale approach, and Streich's application to musical complexity.

The remainder of the paper is structured as follows. Section 2 provides a general formulation of our proposed feature and outlines an efficient implementation. In Section 3 we exemplify the use of the feature with three different basis features and propose a visualisation that summarises the resulting structural change features for a whole track. In Section 5 we provide evidence for the validity of our feature based on a crowd-sourcing experiment. We discuss our approach and future work in Section 6.

2. STRUCTURAL CHANGE ALGORITHM

This section formulates the structural change feature in mathematical terms and provides a description of an efficient implementation.

2.1 Formulation

The formulation of the structural change feature is relatively straight-forward. Since it is designed as a meta-feature, we assume that the m-dimensional audio feature vector $\mathbf{x}_i \in \mathbb{R}^m$, $i = 1, \ldots, N$ has been calculated for all N frames of a music track.

At frame i, the idea is to compare a summary $s_{[i-k+1:i-1]} \in \mathbb{R}^m$ of the features in the k frames to the 'left' to a summary $s_{[i:i+k]} \in \mathbb{R}^m$ of the features in the k frames to the 'right'. [1] For example, in our implementation below the summary is the mean vector.

[1] The dimension of the summary does not have to be the same m as that of the feature, but we use it here for simplicity.

We also assume that we have a non-negative divergence function $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^+$ that assigns a divergence to a pair of feature summaries, for example the Euclidean distance or the Jenson-Shannon divergence (as in our implementation, see Section 3.2). Effectively, d will compare the windows to the left and right of the i^{th} frame.

The characteristic of the structural change feature is that it samples the divergence of the left and right windows at different window sizes w_j, $j = 1, \ldots, n$. The structural change feature at the i^{th} frame is the n-dimensional vector $v_i = \left(v_i^1, \ldots, v_i^n\right)$ of the resulting divergences, where

$$v_i^j = \begin{cases} d(s_{[i-w_j+1:i-1]}, s_{[i:i+w_j]}), \\ \qquad\qquad \text{if } w_j < i < N - w_j + 1 \\ 0 \qquad\qquad \text{otherwise.} \end{cases} \tag{1}$$

While the window widths are arbitrary, it is convenient to think of them as increasing. For example, one possibility is to use window widths increasing by powers of 2:

$$w_j = 2^{j-1}. \tag{2}$$

Using several large windows increases the number of computations, an issue which we address below.

2.2 An efficient implementation strategy

Calculation of the structural change is relatively costly because $2n$ summaries $s_{[:.:]}$ have to be calculated at every frame, two for every window width. Even in the case where the summary is simply the mean of the feature vectors' elements over time computations can become expensive: calculating the sums (required for the means) leads to $2mN \sum_{j=1}^{n}(w_j - 1) = 2mnN(W - 1)$ additions for the whole track, where W is the average window width. For a feature with $m = 12$ dimensions, a track with $N = 2500$ frames, $n = 8$ different window widths and an average window size of $W = 100$ these are nearly 48 million additions. However, when the summary function is indeed the mean, then we can calculate every single summary as just one vector difference (m differences)

$$s[i_1 : i_2] = \mathbf{c}_{i_2} - \mathbf{c}_{i_1} \tag{3}$$

of two vectors from the cumulative feature matrix $C = (\mathbf{c}_0, \ldots, \mathbf{c}_N)$. The matrix C can be easily pre-calculated as

$$\mathbf{c}_i = \sum_{i'=0}^{i} \mathbf{x}_{i'}, \tag{4}$$

where we set $\mathbf{x_0} = \mathbf{0}$. Pre-calculating C is cheap, it costs nN additions, and the additions performed during the structural change calculations are reduced to $2mnN$, i.e. by a factor of W. We have implemented the algorithm in C++ as a library that can be directly included into Vamp feature

plugins [2]. The source code for this library can be obtained from http://github.com/lastfm/.

The window sizes from Equation (2), the mean summary function and the Jenson-Shannon divergence are used in our example implementation below, which represents one particular possibility of configuring the algorithm.

3. IMPLEMENTATION WITH THREE BASIS FEATURES

We apply the structural change algorithm to three different features chosen to represent three qualities of music: chroma (harmony), rhythm and timbre. This section describes the design choices we have made to achieve this.

3.1 The Basis Features

For each of the qualities described by the basis features—chroma, rhythm and timbre—we separately extract the structural change features (SC) as described in Section 2: chroma SC, rhythm SC and timbre SC. All features are extracted from mp3 files sampled at 44100 kHz.

Chroma. Chroma [3] is a 12-dimensional feature of activity values pertaining to the twelve pitch classes (C, C♯, ..., B), a representation of the instantaneous harmony. We use an existing Vamp plugin implementation [3]. The method [6] makes use of the discrete Fourier transform to obtain a spectrogram, maps every spectral frame to the log-frequency space (pitch space) via a linear transform and updates the values to adjust for tuning differences; the chroma vectors are weighted sums of the adjusted pitch space spectral bins. We do not use the approximate transcription (NNLS) step but otherwise use the default parameters with a step size of 11025 samples (250 ms).

Rhythm. The fluctuation patterns (FP) feature [7] was designed to describe the rhythmic signature of musical audio. The FPs are calculated on Hamming-windowed segments of approximately 3 seconds length, with a step size of one second (44100 samples), which are further sub-divided into 256 frames with a length of 512 samples. The main idea is to use the dB amplitude of these 256 frames at different frequency bands as a time series: the spectrum of this time series at a particular frequency band is the FP of that frequency band. We sum the FPs of all frequency bands into one band in order to eliminate timbre influence.

Timbre. The Mel-spectrum is a warped frequency spectrum obtained by taking the discrete Fourier transform of an audio signal, taking the logarithm of the spectral energies to obtain dB values, and mapping the spectrum onto Mel-frequency spaced bins that are linear with respect to human pitch perception. We use 36 Mel-frequency bins. Since the feature is extracted together with the FP, the hop size is one second and the spectral bins are means taken over 256 small frames (512 samples) across a 3 second window.

[2] http://www.vamp-plugins.org/

[3] http://isophonics.net/nnls-chroma

3.2 Window, Summary and Divergence Functions

We choose power-of-two window widths (Equation 2). In order to align time-scales we set $j = 1, \ldots, 6$ for both rhythm and timbre features, and $j = 3, \ldots, 8$ for the chroma feature. This means that the structural change feature is 6-dimensional with window widths (i.e. those of the left or right windows) are 1, 2, 4, ..., 32 seconds.

We use the mean summary function s, which is implemented as described in Section 2.2. Since all basis features can be interpreted as distributions in their respective domains, we normalise each summary vector, and use the Jenson-Shannon divergence as our divergence measure d, i.e. for two normalised summary vectors s_1 and s_2

$$d(s_1, s_2) = \frac{\mathrm{KL}(s_1||M) + \mathrm{KL}(s_2||M)}{2} \tag{5}$$

where $M = \frac{s_1+s_2}{2}$ and KL is the Kullback-Leibler divergence given by

$$\mathrm{KL}(x||y) = \sum_{i=1}^{n} x_i \log(x_i/y_i). \tag{6}$$

3.3 An Example

We have marked a few interesting aspects of the structural change features for the song 'Lucky' in Figure 1 (light colours mean high values). The labels a mark two drum stops, before the first chorus and the first bridge, respectively. Timbre and rhythm SC both show a double bulge, especially in the three bins of short time scales, one at the beginning and one at the end of each drum stop. At b only the timbre SC shows a high value, indicating the beginning of the second chorus (without a clear rhythm change). Label c marks a part with little musical movement: no actual chord changes, but lots of sound variation, including spoken voice excerpts: this is reflected in relatively low chroma SC activity, but relatively high timbre SC activity. Label d marks a calm bridge section (no drums), followed by the key change that leads into the next chorus. Two clear timbre SC peaks show the boundaries of the bridge, and the high chroma long-scale SC values reflect the key change.

4. TRACK-LEVEL SUMMARISATION AND VISUALISATION

In some contexts it is useful to be able to summarise the structural change of a piece of music, for example, summarising the feature for further processing by machine learning algorithms. Summarisation is also necessary to generate track-level visualisations, such as the Audio Flowers, which we present below.

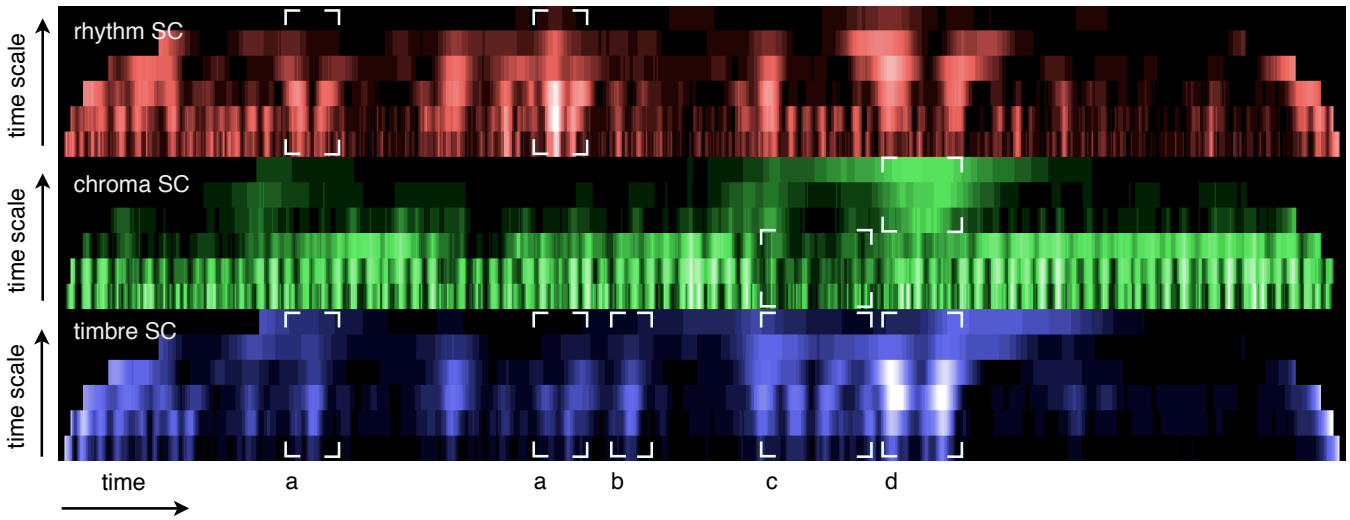


Figure 1: Structural change in the three basis features for the song 'Lucky' as performed by Britney Spears. See Section 3.3.

4.1 Statistics

The most straight-forward way of summarising the SC frames is to take the mean average over all structural change feature frames of the whole piece, resulting in one mean feature vector. In cases where structural change is concentrated in a small part of the piece of music, however, the mean can be misleading because it suggests that the rate of change in the whole piece is relatively high. The median is a more robust average statistic, since it discards such outliers. We use both because mean, median and their difference are interesting properties of a piece of music.

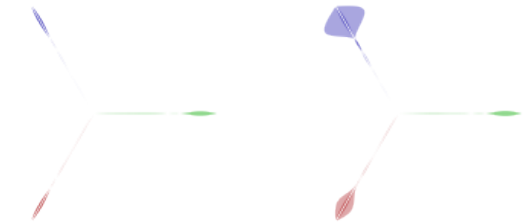
We extracted the structural change features for our three basis features from mp3 files of 17,116 pieces of popular from the British singles charts between 1951 and 2011, then averaged them in two ways by taking the mean and median over time. Since we have six window widths, three basis features and two averages for each of the combinations, each of the tracks has $6 \times 3 \times 2 = 36$ values. For each of the 36 dimensions we apply quantile normalisation (normalised ranking) to spread values within the interval $[0, 1]$ with respect to the whole collection of songs.

4.2 Audio Flowers

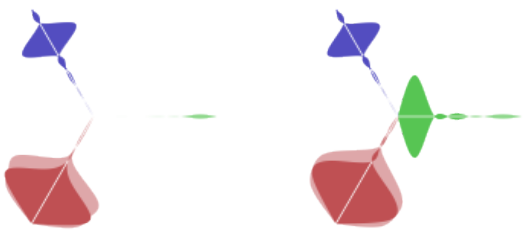
In order to turn the 36 values for each track into an intuitive visual representation (examples in Figure 3), we treat each musical quality separately to create a flower 'petal': red for rhythm, green for harmony, and blue for timbre. In any of the three petals, the central, opaque part visualises the normalised median values, the translucent part corresponds to the normalised mean. The values closest to the centre of the Audio Flower represent short time scales, the values near the tips of the petals represent the longest time scale. The plot is realised by calculating a 100-point smoothed interpolation of the six values. We chose the median to be used for the opaque part because it is a robust average of a track's structural change and is likely to be the most reliable measure. The translucent part is only visible where the mean exceeds the median value. This happens in cases when strong structural changes happen, but on a relatively short section of a track, as we will illustrate below.

Figure 2 shows the results for a few artificially constructed pieces of audio. Figure 2a illustrates 300 seconds of pink noise, Figure 2b 150 seconds of pink noise followed by another 150 of white noise. The white noise Audio Flower shows virtually no sign of structural change, while the Audio Flower of the mixed pink and white noise file has a slight bulge indicating a rare long-term change in timbre (the corresponding rhythm value is slightly raised, too). This indication of 'composedness', or complexity, is exactly what we would expect in that situation (*cf.* Section 1). The other two Audio Flowers are closer to real music: Figure 2c represents a single chord, played on a piano but with two different rhythms alternating at a relativley long time scale of (24 seconds). As we could expect, here too, harmonic change is virtually absent, and the high values towards the tip of the red rhythm petal reflects the long-term rhythm changes. The change in timbre that comes with the rhythm change can be observed, too. Figure 2d was produced from a piece of music with the same rhythm structure, but instead of a single chord we used a cadence, i.e. a more complex chord pattern. The Audio Flower represents this added complexity as high values towards the origin of the green harmony petal, while the rest of the flower remains virtually unchanged.

Figure 3a shows the Audio Flower of the song 'Lucky', which we have already treated in Figure 1. The key change happens only once during the piece, indicated through the high levels of chroma SC at d in Figure 1. Due to this 'out-

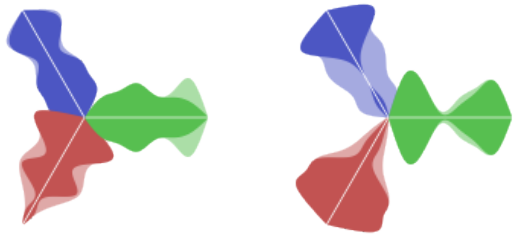


(a) pink noise (b) pink noise, then white noise

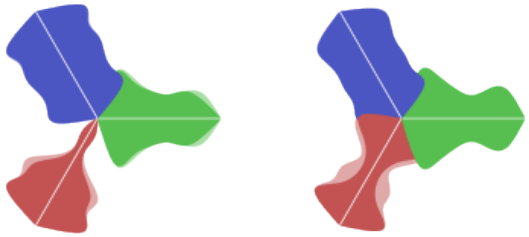


(c) chord, changing rhythm (d) cadence, changing rhythm

Figure 2: Artificial examples: (a) pink noise, (b) pink noise followed by white noise, (c) single major piano chord with different rhythmic sections, (d) repeated major cadences with different rhythmic sections.



(a) Lucky (b) Smells Like Teen Spirit



(c) Time After Time (Lauper) (d) Time After Time (Keating)

Figure 3: Audio Flowers for the songs (a) 'Lucky' (as performed by Britney Spears), (b) 'Smells Like Teen Spirit', and two renditions of 'Time After Time', (c) by Cyndi Lauper, (d) by Ronan Keating.

lier' the normalised median is smaller than the normalised mean at long time scales—the translucent part of the Audio Flower becomes visible.

Figure 3b depicts the Audio Flower of the song 'Smells Like Teen Spirit' as recorded by the band Nirvana. The most striking aspect of this song is the mushroom-shaped timbre petal (blue). This is common in songs that are organised alternating soft and loud sections.

In comparison, the timbre petal of the Audio Flowers in Figures 3c and 3d is decidedly thicker, especially at shorter timescales (towards the origin). In fact, the shape of timbre and chroma petals is very similar between these two Audio Flowers. This is not surprising because they are indeed two renditions of the same song 'Time After Time', one by Cyndi Lauper, one by Ronan Keating. The shape of the rhythm petal is, however, quite dissimilar, which suggests their approaches to rhythm are different. A gallery of further examples can be found at `http://last.fm/playground/demo/complexity`.

5. INTERNET-BASED EXPERIMENT

Finding evidence to support our hypothesis that our features correspond with human perception of structural change is hard because unless the listeners are musicians we cannot assume that they even think in terms of harmony, rhythm or timbre. In order to test whether any correlation can be observed we set up an informal experiment on an Internet page. A participant would randomly be given two 30 second sound excerpts from our collection of chart singles and was then asked to decide which changed more in terms of one of our three basis features. The tracks were chosen to differ in their amount of structural change: the average of the normalised median structural change values [4] for one track was high (> 0.7) and that of the other one was low (< 0.3). The web page clearly states that we look for change and diversity. Upon casting their rating the listener is shown the Audio Flowers of the two songs in question as a reward and is told which of the two our analysis deemed more changeable. The rating was realised as a set of three radio-buttons (first track, second track and a third one labelled 'not sure'). We had no control over whether the participants listened to the tracks before voting.

At the time of writing we have collected 1428 votes from 401 raters with an mean number of 3.9 ratings (median: 2). We analysed the 1165 ratings of the 197 participants who voted at least three times. There is moderate agreement between user ratings and our high and low classes: in 61.4 %

[4] Taking into account the short duration of the excerpts, only the first four dimensions of the features were used in the structural change value.

of all cases users agreed with the automatic analysis. Testing against the null hypothesis of users randomly choosing an answer, we obtain a very low p value of $p < 10^{-14}$, i.e. we are very confident that the participants' choice is not random. This also applies to the three qualities separately: users agree with rhythm SC (60.0%, $p < 10^{-3}$), chroma SC (63.3%. $p < 10^{-6}$) and timbre SC (60.8%, $p < 10^{-4}$).

In all cases the agreement is not very high, but at this stage we can only speculate about the causes: our feature might express something different from what we intended or what participants understood; the un-controlled nature of the experiment may have led participants to randomly choose their rating; the participants may not have had the necessary musical experience to provide meaningful ratings. However, the fact that we found significant agreement for all three features separately suggests that the structural change feature capture musical qualities listeners can relate to.

6. DISCUSSION AND FUTURE WORK

Our implementation presented in Section 3 is only one way of using the structural change feature, and many can be added by using alternatives for the window width function, left/right summary function and divergence function presented here. We are particularly interested in exploring different divergence functions, such as inverse correlation and Euclidean distance (see also [10, Chapter 4]). Using a different divergence function will allow us to use features that are not necessarily non-negative, such as mel-frequency cepstral coefficients (MFCCs) or other chroma mappings.

The proposed feature will allow classic Music Information Retrieval tasks (such as cover song retrieval and genre classification) to access a semantic dimension that is not covered by existing audio features, and hence may lead to improvements in these areas.

Finally, we hope that future studies will reveal how the structural change feature is related to musical complexity as perceived by humans.

7. CONCLUSIONS

We have proposed the novel audio feature *structural change* for the analysis of audio recordings of music. The feature can be regarded as a meta-feature, since it measures the change of an underlying basis feature at different time scales. As part of our proposal we have presented the general algorithm and an efficient implementation strategy of a special case. We have implemented the feature with three different basis features representing chroma, rhythm and timbre. Analysing more than 17,000 tracks of popular music allowed us to find a meaningful normalisation to the feature values. Based on this normalisation we have introduced a track-level visualisation of structural change in chroma, rhythm and timbre. Several of these visualisations, Audio Flowers, have been presented to illustrate the features' characteristics and show that interpreting the amount of structural change as musical complexity is possible. We conducted a informal web-based experiment whose results suggest that our proposed feature correlates with the human perception of change in music.

8. REFERENCES

[1] R. Dawkins. *The blind watchmaker: why the evidence of evolution reveals a universe without design*. Norton, 1996.

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

[3] T. Fujishima. Real time chord recognition of musical sound: a system using Common Lisp Music. In *Proceedings of the International Computer Music Conference (ICMC 1999)*, pages 464–467, 1999.

[4] E. Gómez and J. Bonada. Tonality visualization of polyphonic audio. In *Proceedings of the International Computer Music Conference (ICMC 2005)*, 2005.

[5] F. Heylighen. The growth of structural and functional complexity during evolution. In F. Heylighen, J. Bollen, and A. Riegler, editors, *The evolution of complexity*, pages 17–44. Kluwer Academic, Dordrecht, 1999.

[6] M. Mauch and S. Dixon. Approximate note transcription for the improved identification of difficult chords. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 135–140, 2010.

[7] E. Pampalk, S. Dixon, and G. Widmer. On the evaluation of perceptual similarity measures for music. In *Proceedings of the Sixth International Conference on Digital Audio Effects (DAFx-03)*, pages 7–12, 2003.

[8] R. M. Parry. Musical complexity and top 40 chart performance. Technical report, Georgia Institute of Technology, 2004.

[9] C. Sapp. Harmonic visualizations of tonal music. In *Proceedings of the International Computer Music Conference (ICMC 2001)*, 2001.

[10] S. Streich. *Music Complexity: A Multi-Faceted Description of Audio Content*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain., 2006.