Segmentation and timbre- and rhythm-similarity in Electronic Dance Music

Bruno Rocha[*], Niels Bogaards[+], and Aline Honingh[§]
[*] Faculty of Humanities, University of Amsterdam
[+] Elephantcandy, Amsterdam
[§] Institute for Logic, Language and Computation, University of Amsterdam

Abstract. This report describes the digital humanities project on music similarity. The project is a collaboration between the University of Amsterdam and audio software company Elephantcandy. The project's aim was to investigate timbre and rhythm similarity and to develop an application that finds similar segments of music. In this report three models are described, one for structural segmentation, one for timbre similarity, and one for rhythm similarity of electronic dance music (EDM). The segmentation algorithm performs well on an EDM dataset as well as on a standard MIREX dataset. The timbre similarity algorithm has been tested in a pilot study and preliminary results are presented. Issues related to segmentation and similarity are discussed.

1. Introduction

Similarity in music is a fascinating though complicated concept. Although most people clearly understand when a piece of music is similar to another, a good formalization of the concept of music similarity does not yet exist. Present software applications in this field (recommendation, playlist generation) are often limited to a single outcome and cannot be influenced by the user.

This report is a description of the project `Music Similarity app', based at the Center for Digital Humanities[1]. The project is a collaboration between the University of Amsterdam and Elephantcandy, a company that develops audio applications for mobile devices. The project ran from May 2012 until March 2013. In this project, we have broken down the concept of similarity into sub-similarities, and focus on timbre- and rhythm-similarity, in the restricted domain of electronic dance music (EDM). We have developed an application that has as input a (segment of a) piece of music, and as output a (segment of a) similar piece of music, in accordance with the type of similarity that was specified.

In the academic field of Music Information Retrieval, various systems have been developed that classify music according to a certain type of similarity (Perez-Sancho et al, 2009; Pollastri, and Simoncelli, 2001; Hillewaere et al, 2010; Cilibrasi, Vitanyi, and Wolf, 2004;

[1] http://cdh.uva.nl/

Wiggins, 2007; Chew, Volk, and Lee 2005; Honingh and Bod, 2011). On the other side, in industry, a number of tools have been released that can recommend similar music (Apple Genius, last.fm, Pandora). Such systems and tools, however, often (1) rely on metadata and not on the actual audio, (2) consider similarity as a holistic entity, and (3) consider only complete musical records. As a result, only limited functionality can be provided to the end user.

Let us briefly go through the shortcomings of existing systems.

Most existing music-recommendation systems use metadata (keywords tagged by the user which can include information about artist, title, genre and more), meaning that the music itself (the audio file) is not studied (Lamere 2008). Therefore, the amount of music that can be used as both input and output is limited, and the functionality is limited to finding matches that have the same label. This project focuses on content-based music information retrieval, in which the audio is studied. In this way, we can use any kind of music, and have access to all musical information contained in the audio.

Musical similarity consists of many facets, for example, tempo, rhythm, meter, instrumentation and pitch contour. Current research and industrial tools often treat similarity as a monodimensional property, aiming for an arbitrary 'best match'. However, as will be argued later, similarity depends on context and it is therefore useful to expand the notion of similarity into sub-similarities.

Most studies in the area of music similarity concentrate on the similarity of pieces of music or songs as a whole. We can, however, imagine that a piece of music is similar to only a part of another piece of music, for example its introduction. The overall similarity between the two pieces will be therefore not that high, while the similarity between the first song and the introduction of the second could be of great importance. Therefore, in this project, we will focus on the similarity of *segments* of music.

Since the topic of music similarity, even when restricted to timbre and rhythm similarity of music segments, remains a broad subject, we decided to treat it in the restricted domain of electronic dance music (EDM). The choice for this genre was motivated by the collaboration with audio software company Elephantcandy, which identified a specific need for similarity tools in this genre.

In this report, we will describe the different parts of the project. We start with an introduction in electronic dance music, the chosen domain of our application. In section 3, we will describe the segmentation algorithm, and the evaluation thereof. Section 4 describes the similarity algorithms for both timbre and rhythm. We end with concluding remarks and discussion in section 5.

2. Electronic Dance Music

Electronic Dance Music (EDM) is a label that defines a metagenre encompassing a heterogeneous group of musics made with computers and electronic instruments (McLeod, 2001). Most EDM tracks are made with the expectation of being combined with other tracks and danced to. However, some genres, although drawing on the conventions of EDM, are not suitable for the dance floor or written intentionally for not dancing (Butler, 2006).

EDM was until recently (with some sporadic exceptions) an underground culture, i.e. cultivated outside the view of the general public eye (Fikentscher, 2000), but it has risen to the mainstream charts of the music industry (Greenburg, 2012). Today it has become common for established Top 40 artists and producers to infuse elements of popular EDM styles in their music. EDM "has broken free from the underground to become the driving beat behind pop music and product sales, the soundtrack of choice for a new generation" (Ferguson, 2012). In 2005, the Grammy Awards added a Best Electronic/Dance album category. In 2012, the American Music Awards added a Favourite Electronic Dance Music category.

Almost all EDM share certain musical characteristics: (1) steady tempo, mostly in the range of 120-150 BPM (dependent of genre); (2) a repeating bass drum pattern is almost always present (Butler, 2006).

EDM has numerous subgenres, like for example house, techno, trance or drum 'n' bass. The subgenres can usually be divided into two different styles: "four-on-the-floor" and "breakbeat". Each style is concerned with a specific type of rhythm. However, these two different rhythms share certain aspects of design: (1) they are presented in cycles; (2) the duration of a cycle is almost always duple, i.e. rhythms occur in groups of 2, 4, 8, 16, 32, and 64 (Butler, 2006).

Interviews with EDM specialists, conducted by Yenigun (2012) and Ryce (2012) revealed some new perspectives on EDM. For example, it was said that EDM can be divided into two tendencies: (1) a "trackier" tendency, where the focus is on the textures and rhythms, not on the melody - in the sense that there are no memorable riffs or hum-along-to tunes; (2) a "song oriented" tendency, with a structure resembling pop where vocals dominate. It has furthermore been suggested in these interviews that the focus on melodies and vocals turns EDM "accessible to people that may not have liked the looping styles of deeper stuff".

2.1. Structure Analysis in EDM

The fundamental unit of musical structure in EDM is the "loop". As a fundamental structural idea of EDM, the cyclical repetition of rhythmic patterns manifests itself on multiple levels: not only in loops, but also in sequences, and ultimately within the structure of a complete track, as embodied in the form of a continuously revolving record (Butler, 2006).

Particular instruments show characteristic rhythm patterns that shape the rhythmic and metrical profile of a track. It is often the bass drum pattern that listeners refer to when they describe "the beat" in EDM. Playing with the beat is essential to the metrical, textural, and formal processes that occur in EDM. The most common phenomenon is the removal of the bass drum - followed by its eventual return. The dynamic of removal and return is pervasive within EDM, appearing at some point in nearly every track (Butler, 2006).

Timbre, often also referred to as 'texture', also stands out as a primary compositional parameter in EDM. Yeston (1976, as cited in Butler, 2006) stated that timbre is "the criterion by which rhythmic sub-patterns may be differentiated most easily". Most of the timbral changes that occur in EDM involve an element either entering or leaving the mix. In Butler (2006), DJ Shiva and Stanley described a prototypical structure of EDM tracks. They based their descriptions mainly on timbral changes.

Butler (2006) explains how EDM producers rely on sequences (for example a sequence of notes, possibly consisting of multiple voices) to create a track: they use only one or two sequences through the course of an entire track, creating textural variety by muting or unmuting selected parts as the sequence repeats; they form the track from a succession of many different sequences. These sequences usually consist of four or multiples of four measures of four beats. As the DJs Butler interviewed stated, in EDM "everything happens in four", be it beats, measures, or hypermeasures. However, empirical analysis in the current project showed it has become increasingly common for producers to introduce an element of surprise, typically by adding one measure at the end of some segments.

3. Unsupervised Detection of Structural Changes in EDM

The segmentation of time series into meaningful, coherent units by automatically detecting their boundaries is a challenge crossing several scientific domains (Serrà et al, 2012). A musical segment is a region with some internal similarity or consistency in a given feature space, such as timbre or instrumentation, implying that it has temporal boundaries at its start and end (Casey et al, 2008). Tzanetakis and Cook (2000) stress the importance of segmentation in MIR, where it is better to consider a song as a collection of distinct regions than as a whole with mixed statistics. Performance in audio similarity can benefit from segmenting the tracks beforehand (Casey et al, 2008).

As pointed out in section 2.1, timbral changes are essential for EDM producers when considering structural changes. Aucouturier and Sandler (2001) argue that, to segment a song into its relevant sections, one should discard any pitch and harmonic information and focus only on timbre. Thus, to allow an efficient segmentation of the audio data, the same authors define how the ideal feature set should be: (1) a perceptually realistic measure of the similarity of timbres - similar textures must be represented by close "points" in the multi-dimensional feature space, and the other way round; (2) relatively independent of pitch, as we don't want to segment the different notes or events within a single texture.

Most recent algorithms for music structure segmentation use a combination of extracted features (usually chroma vectors and MFCCs), segmentation methods (which, following Paulus et al. (2010), can be divided into three main categories: repetition-based, novelty-based, and homogeneity-based), and labelling/grouping techniques. We do not need the latter step, since, as explained in section 2.1, EDM tracks can be formed by juxtaposition of several different sequences. When this happens there is no repetition of segments in a track. Therefore, here we only need to (1) extract features, and (2) divide the music into segments, based on these features. In order to take into account the dynamic evolution of a feature, the analysis has to be carried out on a short-term window that moves chronologically along the temporal signal; each position of the window is called a frame (Lartillot and Toiviainen, 2007). After extracting the relevant features on subsequent frames one has to calculate the distance between each frame and all the others, according to a certain distance measure. The largest calculated distances represent the segment boundaries. Once the algorithm estimates the boundaries, each segment in a track will be represented by a set of relevant features and their statistical analysis.

Most of the algorithms run roughly real-time. Therefore, any very large-scale effort to automatically segment music audio will require significant computational resources (Ehmann et al, 2011). We will explain all steps of the segmentation algorithm below.

3.1. Downsampling

We start by loading the audio file and downsampling it. We downsample the signal by a factor of four, reducing the sampling rate from 44100 Hz to 11025 Hz, thus reducing the size of the data four times. We do it for a practical reason, as it helps to run the algorithm faster. Systematic tests on the audio files show that the results of the novelty detection (explained in section 3.6) do not change much up to a factor of four downsampling, which makes sense as most of the spectral information can be found below 5000 Hz. The sampling rate must be higher than two times this value to make sure the Nyquist-Shannon sampling theorem criterion is maintained (Shannon, 1949).

3.2. Detection of first bass drum downbeat

Many EDM tracks begin with beatless intros and culminate in “turning the beat around”, a phenomenon that happens when people perceive a certain metrical structure that is violated (usually by introducing a beat on the perceived off-beat) (Butler, 2006). For this reason, the entrance of the bass drum in an EDM track often results in a decisive metrical representation (Butler, 2006). In some cases, DJs may even skip beatless intros and start playing from the first bass drum beat, representing the start of the main structure of the track, which makes its detection a critical step for the performance of the segmentation algorithm.

To detect the first bass drum downbeat, we start by applying a bandpass filter between 50 and 150 Hz, which is the region where most of the energy of the kick-drum is usually found. We then compute the global energy of the filtered signal by taking the root average of the square of the amplitude, also called Root Mean Square (RMS), on non-overlapping windows of 30 seconds, in order to find in which part of the audio file is the beat likely to start (beatless intros usually have low-energy in the low-frequency region). An onset detection is then performed on the thirty seconds window where the energy rises abruptly, leaving us with candidates for the first downbeat. We select the first that exceeds a given threshold and save the previous part as the first segment.

Figure 1 shows the audio waveform after downsampling and before filtering. Bandpass filtering between 50 and 150 Hz leads to the waveform depicted in Figure 2. Figure 3 portrays the RMS energy curve on windows of 30 seconds. The onset curve is reproduced in Figure 4.

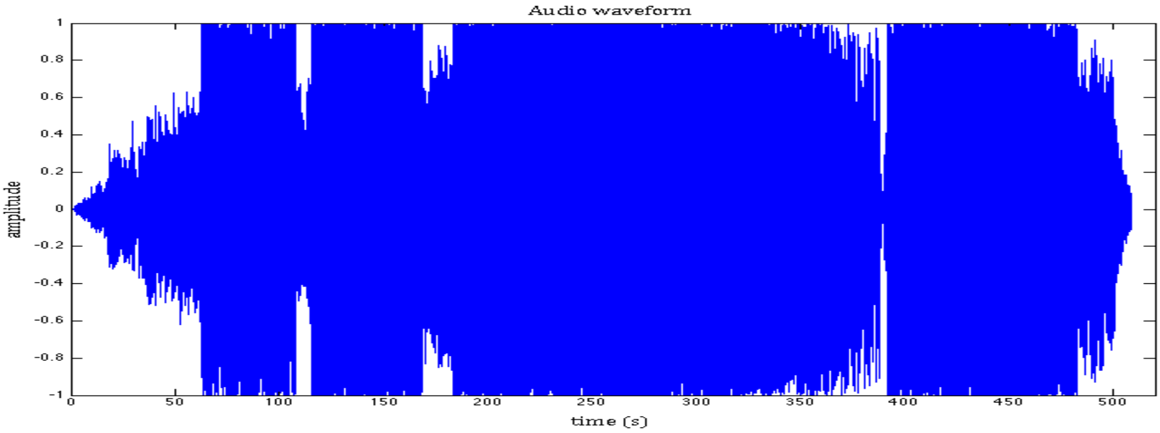


Figure 1: Audio waveform after downsampling

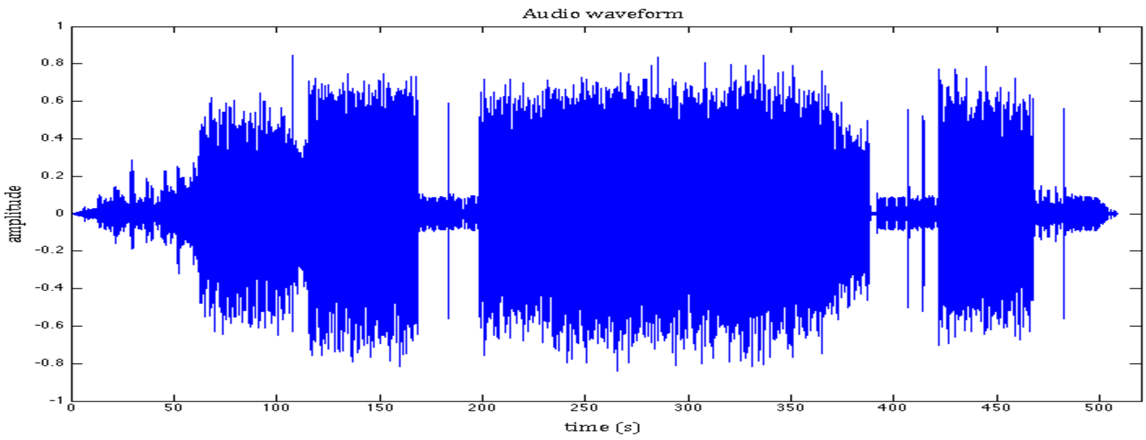


Figure 2: Audio waveform after bandpass filtering

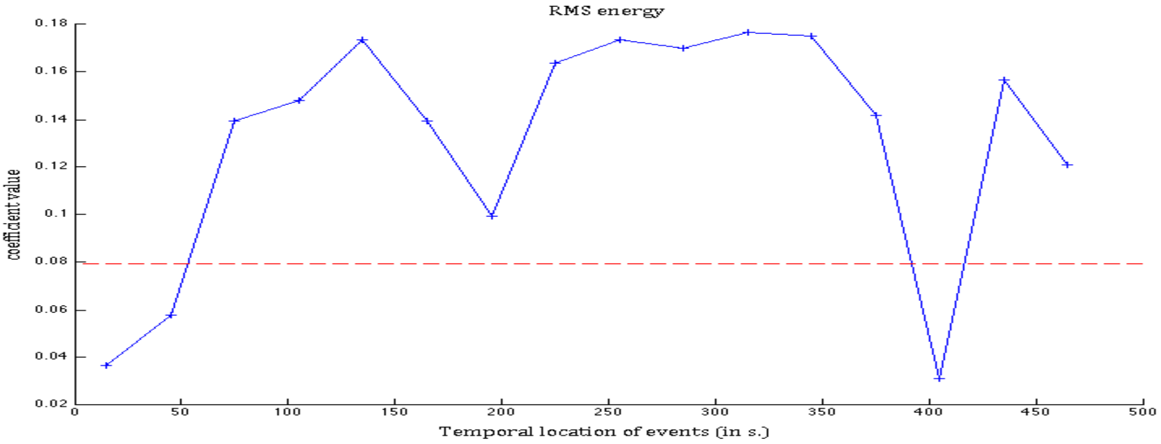


Figure 3: RMS energy curve; red line corresponds to the threshold; in this track, the threshold is exceeded in the third instance, which means that the onset detection will be performed between 40 and 70 seconds

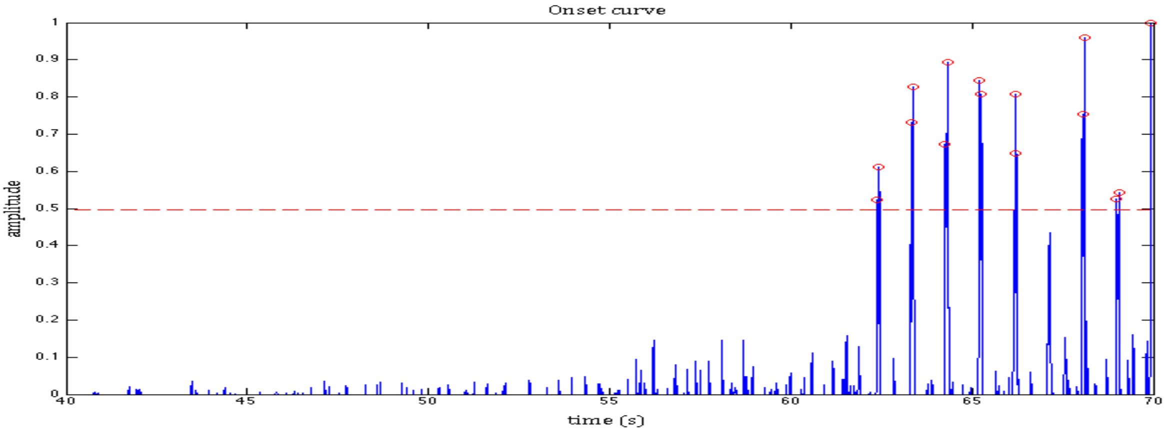


Figure 4: Onset detection on 30 seconds window; red circles indicate detected onsets

3.3. Tempo estimation and confidence measure

We perform tempo estimation in order to detect the duration of a beat. This is important because: (1) all features (for both the segmentation and the similarity tasks) are extracted on beat-related frame lengths; (2) musically informed rules (section 3.7) rely on the beat duration to improve the accuracy of the boundary estimation.

Looking at local correlation between samples we can evaluate periodicities in a signal. This is called an autocorrelation function and it is obtained by multiplying point per point the signal with a shifted version of itself. When the shift difference corresponds to a period of the signal, the summation of both gives a very high value, as the two signals are highly correlated. An autocorrelation function is computed on the onset detection curve and translated into the frequency domain in order to be compared to a spectral decomposition of the onset detection curve, and the two curves are subsequently multiplied (Lartillot and Toiviainen, 2007). The result is a curve with peaks as indications of the most predominant periodicities found in the track. We then perform peak picking and select the highest peaks above a certain threshold. The highest peak is always selected as the tempo of the track.

A binary confidence measure – telling us how certain the algorithm is that the detected tempo is correct – is then derived from the harmonic relation between the found peaks. When only one peak is detected or all the observed peaks are harmonically spaced (which would give alternative tempos that are for example two or three times as fast), the estimated confidence value is 1. It is not a problem if the detected tempo is in a harmonic relation with the 'real' tempo, since this would just refer to another tempo-level that can also be perceived (in the same way as one can tap along with music on different tempo-levels). If there are several peaks with no harmonic relation between the spacing of the peaks, the estimated confidence value is 0. This measure determines whether the musically informed rules (section 3.7) are applied. Figures 5 and 6 show examples of tempo estimation.

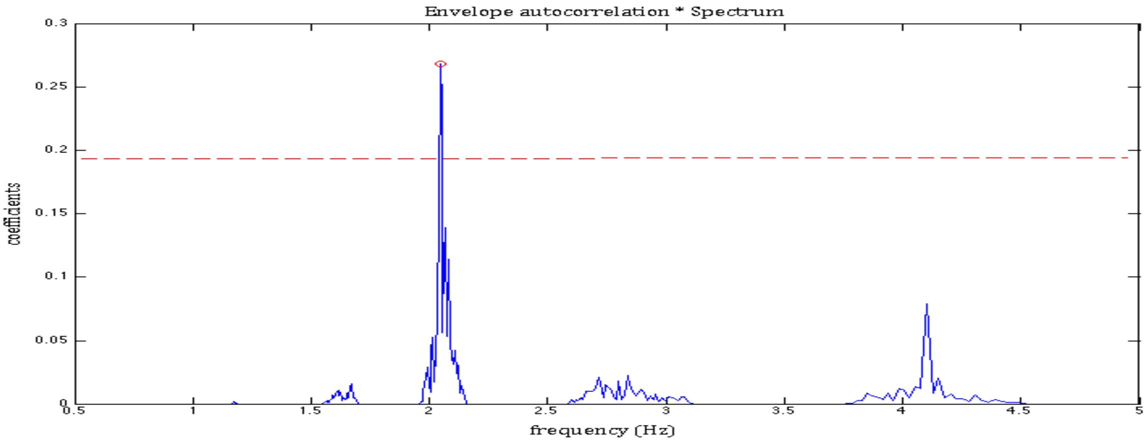


Figure 5: Tempo estimation with confidence = 1; as there is only one peak above the threshold, this is selected as the tempo of the track; the musically informed rules (explained in section 3.7) are applied

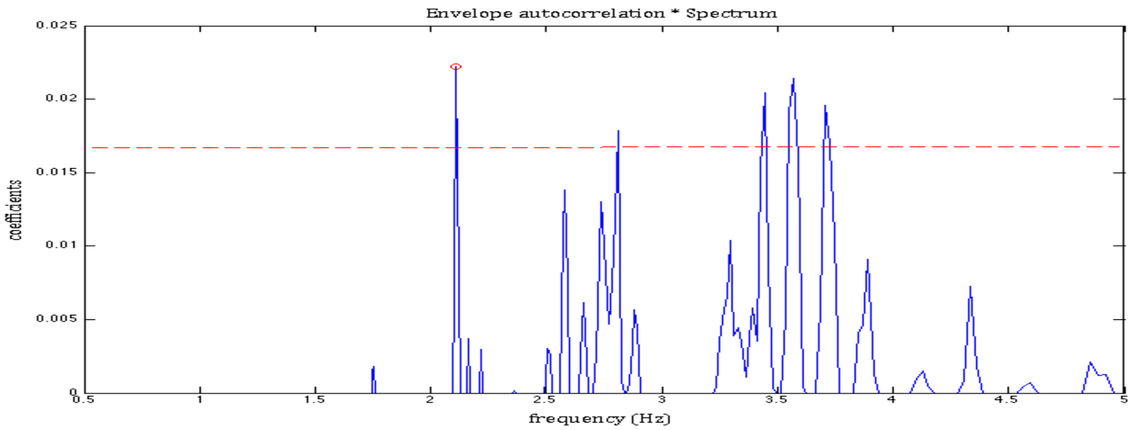


Figure 6: Tempo estimation with confidence = 0; as there are five peaks above the threshold and no harmonic relation between the peaks is found, the highest peak (indicated by the red circle) is selected as the tempo of the track, but the musically informed rules (explained in section 3.7) are not applied

3.4. Magnitude spectrum

After having the tempo score in beats per minute (BPM) and building a vector with all the probable beat positions, we compute the magnitude spectrum of each frame of the signal. This is a decomposition of the energy of the signal along frequencies and it can be performed using a Fast Fourier Transform (FFT). The frames are beat-aligned with 87.5% overlap so that we decompose the energy along frequencies for each beat of the track. To solve problems deriving from performing Fourier transforms on finite signals, generally a windowing technique is used; in our case, the Hamming window was chosen (Smith, 2011).

3.5. Logarithm cepstrum

Bogert, Healy, and Tukey (1963, as cited in Oppenheim and Schafer, 2004) termed the spectrum of the log of the spectrum a *cepstrum*. They chose to coin this term to avoid confusion while emphasizing connections to similar concepts. The domain represented in this operation is neither the frequency nor the time domain, but what they called the *quefrency* domain.

To allow an additive separability of product components of the original spectrum a natural logarithm is performed on the magnitude spectrum. We then compute a FFT of the magnitude spectrum in order to convert from the frequency domain back to the time domain to find periodic sequences in the signal. We end up with a cepstrum and the results can be expressed in the quefrency domain.

3.6. Novelty detection

We then compute the cosine distance between each possible pair of frames from the cepstrum data to get a self-similarity matrix, which can be seen in figure 7. Convoluting along the main diagonal of the similarity matrix using a Gaussian checkerboard kernel[2] yields a unidimensional linearly normalized novelty curve that indicates the temporal locations of significant textural changes. We use a kernel size of 128 frames, corresponding to approximately 30 seconds.

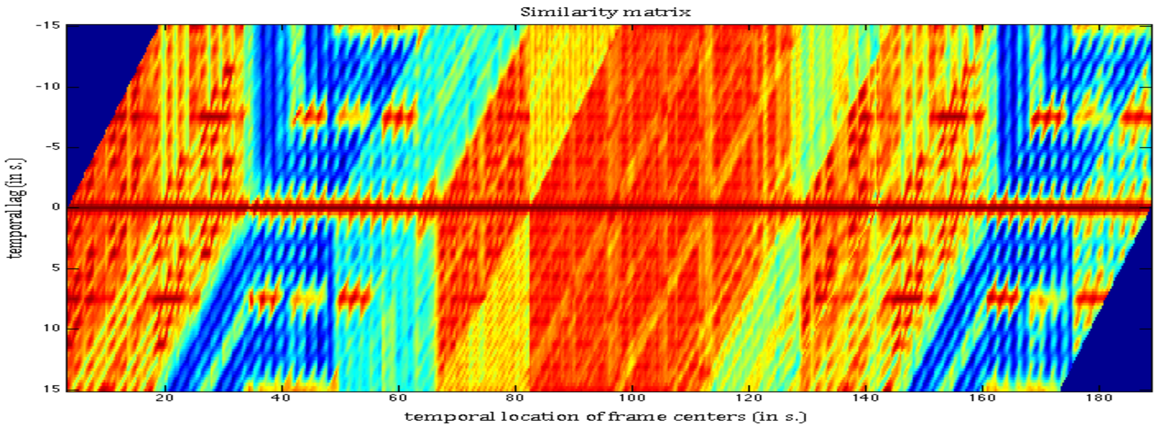


Figure 7: Similarity matrix with kernel size of approximately 30 seconds

Figure 8 shows the novelty curve. Positions corresponding to the above-threshold-peaks of this curve are selected as segment boundaries.

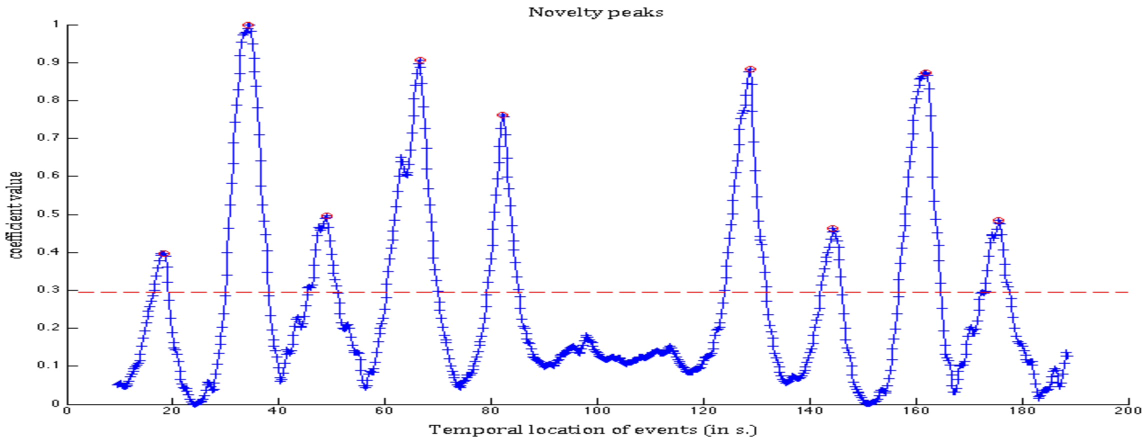


Figure 8: Novelty curve

3.7. Musically informed rules

Butler (2006) categorizes sounds in EDM as “rhythmic”, “articulative”, or “atmospheric”. For the purpose of segmentation, articulative sounds, which are brief and intermittent, are very important. They usually appear before structural boundaries, such as the beginning of a measure or multimeasure group, in order to raise expectation for a segment boundary for the listener. Besides the removal or addition of the bass line or the bass drum, the use of articulative sounds is the most effective way to demarcate segments.

[2] A kernel is a convolution matrix that is useful for several signal processing methods. In this case it consists of the diagonal with a certain width of a similarity matrix.

As the novelty detection is based on textural changes and the timbres of articulative sounds are frequently quite distinct from the neighbours', novelty peaks are detected when these sounds occur. However, the relevant structural changes we want to detect are usually synchronous with the beginning of the aforementioned sequences (section 2.1).

To overcome this displacement, we propose a set of heuristic rules to align the obtained novelty peaks with the most probable to be perceived by listeners - for the tracks on which the tempo was estimated with confidence. We analyze the distances between peaks and update them at each iteration, forming a dynamic structure.

Furthermore, to account for the extra measure issue (explained in section 2.1), an asymmetric weight was applied, such that the gravitation toward the 8th or 16th measure mark is stronger when a boundary is detected before than when it is detected after that mark. Figure 9 shows the effect of the rules on a hypothetic track.

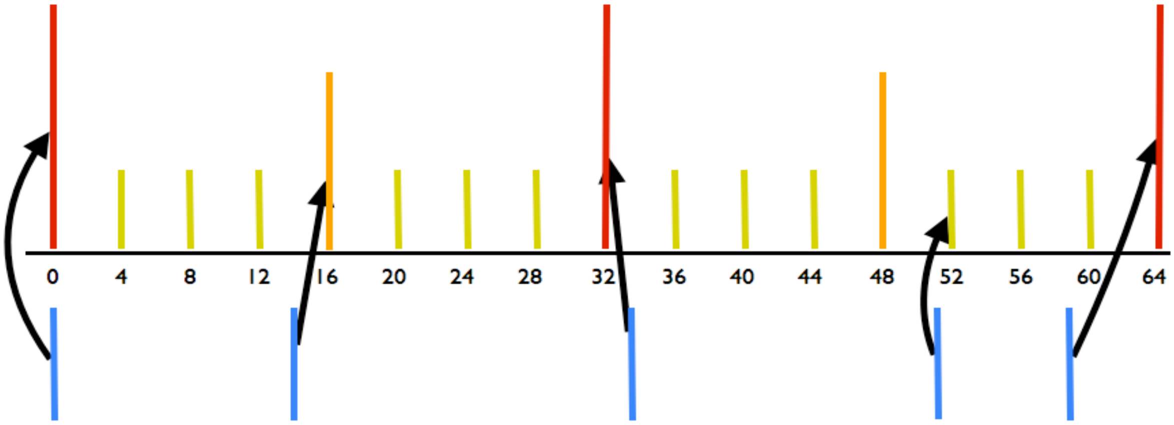


Figure 9: Application of musically informed rules to detected boundaries. Timeline is shown in beats (0 corresponds to the first detected beat; 4 beats = 1 measure). Heuristic rules dictate a dynamic and asymmetric weight towards the 8th and 16th measures.

For the tracks that had a tempo estimation with confidence=0, the detected boundaries remain unchanged, as the changes would most probably result in a less precise estimation of the segment boundaries. However, for the tested datasets, more than 90% of the tracks had confidence=1.

3.8. Evaluation

We have implemented an algorithm for the detection of structural changes in EDM. We employed a novelty-based method with the help of cepstrum features in order to get the best possible performance.

Table 1 shows the results obtained for different datasets using the same parameter settings. The datasets that have been used are: (1) EDM[3], an in-house dataset specially created for this project, consisting of 35 tracks - annotated by the authors - from 19 artists; (2) RWC Pop,

[3] The annotations are available for research purposes on request to the authors.

created by Goto et al. (2002), annotated by two groups of researchers – RWO corresponds to the annotations of the dataset creators and RWQ corresponds to the annotations that Bimbot et al. (2010) did for the Quaero[4] project; (3) Eurovision dataset, annotated by Bimbot et al. (2011). Found segment boundaries are considered correct if they are within ±0.5 seconds or ±3 seconds from a border in the ground truth annotations. Based on the matched hits, *boundary retrieval recall rate*, *boundary retrieval precision rate*, and *boundary retrieval F-measure* are calculated. These are some of the evaluation measures used in MIREX[5], an annual evaluation contest for MIR algorithms.
Precision and recall are defined as:

$$precision = \frac{tp}{tp + fp} \qquad \textbf{Eq. (1)}$$

$$recall = \frac{tp}{tp + fn} \qquad \textbf{Eq. (2)}$$

In these equations, *tp*= `true positives', the number of correctly identified segment boundaries, *fp* = `false positives', the number of indicated segment boundaries that do not correspond to true segment boundaries, and *fn*= `false negatives', the number of segment boundaries that have not been identified by the algorithm.

The F-score is given by:

$$f = 2 \cdot \frac{precision \cdot recall}{precision + recall} \qquad \textbf{Eq. (3)}$$

The algorithm performs well on the EDM dataset. As can be seen from table 1, the musically informed rules increased the F-score with around 10 points on the 0.5s tolerance-window level. Although this method was created specifically for EDM, results on the RWC Pop dataset would be in the top 3 of best performing algorithms submitted to MIREX 2012, with its best performing algorithm having F(3s) = 0.77 on RWQ and F(3s) = 0.71 on RWO. This suggests that structural changes in pop music might have the same periodicity as in EDM. This method performs poorly on the Eurovision dataset. An explanation for this might be that, in this song contest, pop music is usually mixed with traditional music from several European countries of which the structural boundaries may be quite distinct.

[4] http://www.quaero.org

[5] http://www.music-ir.org/mirex/wiki/2010:Structural_Segmentation

Dataset	P0.5s	R0.5s	F0.5s	P3s	R3s	F3s
EDM no rules applied	37.10	51.48	41.63	63.62	86.34	70.80
EDM rules applied	46.52	62.87	51.67	62.15	84.83	69.38
RWO no rules applied	30.12	27.07	27.81	70.71	64.64	65.81
RWO rules applied	28.10	23.95	25.28	70.11	63.70	65.08
RWQ no rules applied	27.58	25.67	26.05	67.48	62.01	63.28
RWQ rules applied	31.40	27.86	28.99	66.74	61.25	62.59
EUR no rules applied	8.80	8.59	8.37	43.55	42.93	41.92
EUR rules applied	9.27	9.15	8.86	43.85	43.55	42.39

Table 1: Boundary retrieval precision rate (P), recall rate (R) and F-score (F) with two tolerance windows: ±0.5 seconds and ±3 seconds. Three annotated datasets were used: in-house (EDM), RWC (original (RWO) and Quaero (RWQ) annotations) and Eurovision (EUR)

3.9. Case Study: Basement Jaxx – "Red Alert"

This track from Basement Jaxx shows that, when the first downbeat is correctly detected and the tempo well estimated (see Table 2), applying the musical rules could lead the segmentation algorithm to achieve the same performance on the ±0.5 seconds tolerance measures as it does on the ±3 seconds measures (see Table 3).

Red Alert	First Downbeat	Tempo (bpm)
Annotated	7.9686	127[6]
Estimated	8.0599	126.823

Table 2: Annotated/estimated first downbeat/tempo for the song "Red Alert" by Basement Jaxx

Red Alert	P0.5s	R0.5s	F0.5s	P3s	R3s	F3s
no rules applied	40.00	28.57	33.33	93.33	66.67	77.78
rules applied	93.33	66.67	77.78	93.33	66.67	77.78

Table 3: Boundary retrieval precision rate (P), recall rate (R) and F-score (F) with two tolerance windows (±0.5 seconds and ±3 seconds) for the song "Red Alert" by Basement Jaxx

[6] Annotation extracted from *Beatport*, an online music store specializing in EDM.

4. Similarity

When are two things similar? Similarity is one of the most central theoretical constructs in psychology. An important Gestalt principle of perceptual organization is that similar things will tend to be grouped together (Medin, Goldstone, and Gentner, 1993), and similarity plays a crucial role in making predictions because similar things usually behave similarly (Goldstone and Son, 2005).

Several models for similarity have been proposed. These models have had an impact on fields such as statistics, pattern recognition, or data mining, and are usually divided into four categories: geometrical (Shepard, 1962a, 1962b), feature based (Tversky, 1977), alignment based, and transformational – a comprehensive overview of the models can be found in Goldstone and Son (2005).

4.1. Music Similarity

In music information retrieval (MIR), the literature is rich in music similarity functions and algorithms (Pampalk, Dixon, and Widmer, 2003; Aucouturier and Pachet, 2004a; Aucouturier, Pachet, and Sandler, 2005; Berenzweig et al., 2004; Bogdanov et al., 2011). However, there is no comprehensive approach to similarity in the domain of music, thus the challenge of relating musical features to the listeners' concept of similarity is a major problem in MIR (Volk, de Haas, and van Kranenburg, 2012).

The starting point of this project's method for timbre similarity is the combination of continuous dimensions and discrete features in a single model, as proposed by Navarro and Lee (2003). The method for rhythm similarity, besides dimensions and features, incorporates alignment (i.e. structure relationships between features; the order of the objects being compared is taken into account).

Goldstone and Son (2005) describe two types of dimensions: the ones that are described as being more or less (e.g. loud is more sound than soft), which can be represented by sequences of nested feature sets; the ones defined by qualitative attributes (e.g. spectral shape), which can be represented by chains of features (imagine three polyphonic textures composed of two instruments each: (1) *guitar* and *piano*, (2) *guitar* and *drums*, (3) *bass* and *drums*; if an imaginary axis can be drawn in which (2) lies between the other two textures, then this can be featurally represented if (1) and (2) share features that (2) and (3) do not share).

4.2. Timbre Similarity

Studies in timbre perception have historically yielded results indicating that the phenomenon of timbre is multidimensional, with a number of factors interacting to produce the exact tone quality that is perceived by a listener (Fales, 2004). These factors have been identified to include, among others, spectral flux, spectral centroid, and attack time (McAdams et al, 1995; Burgoyne and McAdams, 2008; Peeters et al, 2011). These studies focused on monophonic timbres. On the contrary, here we want to describe polyphonic textures, which the aforementioned features cannot fully represent.

For our purposes, we have empirically made a selection of a small number of features to describe the timbre in EDM. We will now describe the three types of features that we believe capture the most relevant dimensions of a polyphonic texture for comparison with other textures.

a) Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) are extensively used in MIR algorithms to represent the spectral envelope of a given sound, which is one of the most salient components of timbre. We calculate them by first computing the power spectrum successively on frames with the duration of a beat, followed by logarithmically positioning the frequency bands on the Mel scale, and finally performing a discrete cosine transform on the bands (Lartillot and Toiviainen, 2007).

MFCCs fall into the second category of dimensions referred by Goldstone and Son (2005), as it is impossible to set a hierarchy of spectral envelope. The number of MFCCs that well represent a spectral envelope is a matter of great discussion. The low order MFCCs account for the slowly changing spectral envelope, while the higher order ones describe the fast variations of the spectrum (Aucouturier and Pachet, 2002). Therefore, while it is true that the more MFCCs we compute, the more precise the approximation of the signal's spectrum is, a large number of MFCCs may not be appropriate, as we are only interested in the spectral envelope and not in the finer details of the spectrum (Aucouturier, Pachet, and Sandler, 2005). The same authors reported an ideal value of 20 coefficients, which we implemented.

For the computation of these features, we frame the signal into half-overlapping windows with duration of a beat and calculate the mean of each coefficient for each segment, ending up with twenty values per segment.

b) Spectral Flatness

Facing the problem of audio matching (i.e. finding in a database the audio that matches a given example), Herre, Allamanche, and Hellmuth (2001) searched for features that, while being perceptually meaningful, are independent of absolute level and coarse spectral envelope. This led the authors to examine features relating to the tonal character (the notion of *tonality* as used in the perceptual audio coding field, cf. Hellman, 1972) of the signal within particular frequency bands.

As known from coding theory, the maximum gain that can be recovered by redundancy reduction using predictive coding methods or transform coding is determined by the flatness of the signal's power spectral density and is related to the so-called Spectral Flatness (Jayant and Noll, 1984, as cited in Herre et al., 2001). Spectral flatness measures the *sinusoidality* of a spectrum (Peeters, 2004). It indicates whether the distribution of the spectrum is smooth or spiky, and results from the simple ratio between the geometric mean and the arithmetic mean (Lartillot and Toiviainen, 2007):

$$\frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)}$$ **Eq. (4)**

For the computation of these features, we first split the spectrum in four bands[7] - 20 to 200 Hz, 200 to 800 Hz, 800 to 3200 Hz, and 3200 to 5512.5 Hz (half of the sampling rate, 11025 Hz). Then we frame the signal into half-overlapping windows with the duration of a beat. Finally we calculate the mean spectral flatness for each band, ending up with four values per segment.

c) Dirtiness

Helmholtz (1863/1954) introduced the term “auditory roughness”, also referred to as sensory dissonance, in the psychoacoustics literature. It is related to the beating phenomenon that occurs whenever a pair of sinusoids is close in frequency in a short period of time (Plomp and Levelt, 1965). It can be considered as an attribute of timbre, as it is usually described as a function of a signal’s amplitude envelope and corresponding spectral distribution (Vassilakis and Kendall, 2010).

Sethares (1998) proposed a method for the estimation of roughness. For each pair of spectral peaks, the corresponding elementary roughness is obtained by multiplying the two peak amplitudes altogether, and by weighting the results with the corresponding factor given on the dissonance curve. The summation of all these values is the total roughness (Sethares, 1998). Vassilakis (2001) developed a variant of Sethares’ model with a more complex weighting, adding a term to the equation that accounts more reliably for the dependence of roughness on sound pressure level and amplitude fluctuation (Vassilakis, 2001, Eq. 6.23).

We took the notion of roughness and approached it from a different perspective. “Dirtiness” is a concept applied by EDM listeners and producers when referring to a particular sound quality that is pervasive in EDM synthesizers – there is even a subgenre of EDM called “Dirty Dutch” (“Styles of House Music”, 2013) and numerous online videos teach how to achieve a “dirty” synth[8] sound.

This phenomenon of dirtiness in EDM is parallel to ''heaviness'' in heavy metal as heavy metal listeners use the concept of “heaviness” to describe a range of instrumental timbres, particularly guitars (Fales, 2004). Both dirtiness and heaviness are concepts born from the musicians and fans. We tried to use Fales approach of linking a verbal description of a tone quality to acoustic features, rather than searching for agreement among listeners that a specific sound is characterized by a given descriptor. Spectral analysis revealed that dirtiness might be partly explained by the detuning that producers apply to their synth sounds. This detuning is characterized by a varying stream of frequencies very close to the harmonics of the fundamental frequency we perceive as the pitch of the played sound, which can therefore be described using the concept of roughness.

[7] These four bands are the same for *dirtiness*, *event density,* and *onset patterns*

[8] Synth is a common abbreviation of synthesizer.

We are not interested in the value of roughness at each instance but in its value over a larger period of time with a very high frequency resolution. For this reason, we compute roughness values in half-overlapping windows of 8 beats. For the computation of roughness we use Vassilakis' (2010) model, mentioned before. Dividing the spectrum into four bands, we then calculate the mean for each band, ending up with four values per segment.

The aforementioned features (20 MFCCs, 4 Spectral Flatness values, and 4 Dirtiness values) together make a feature vector that describes the timbre of a segment. The similarity between two different timbres is then described by calculating the Euclidian distance between the two associated feature vectors.

4.3. Evaluation of Timbre Similarity

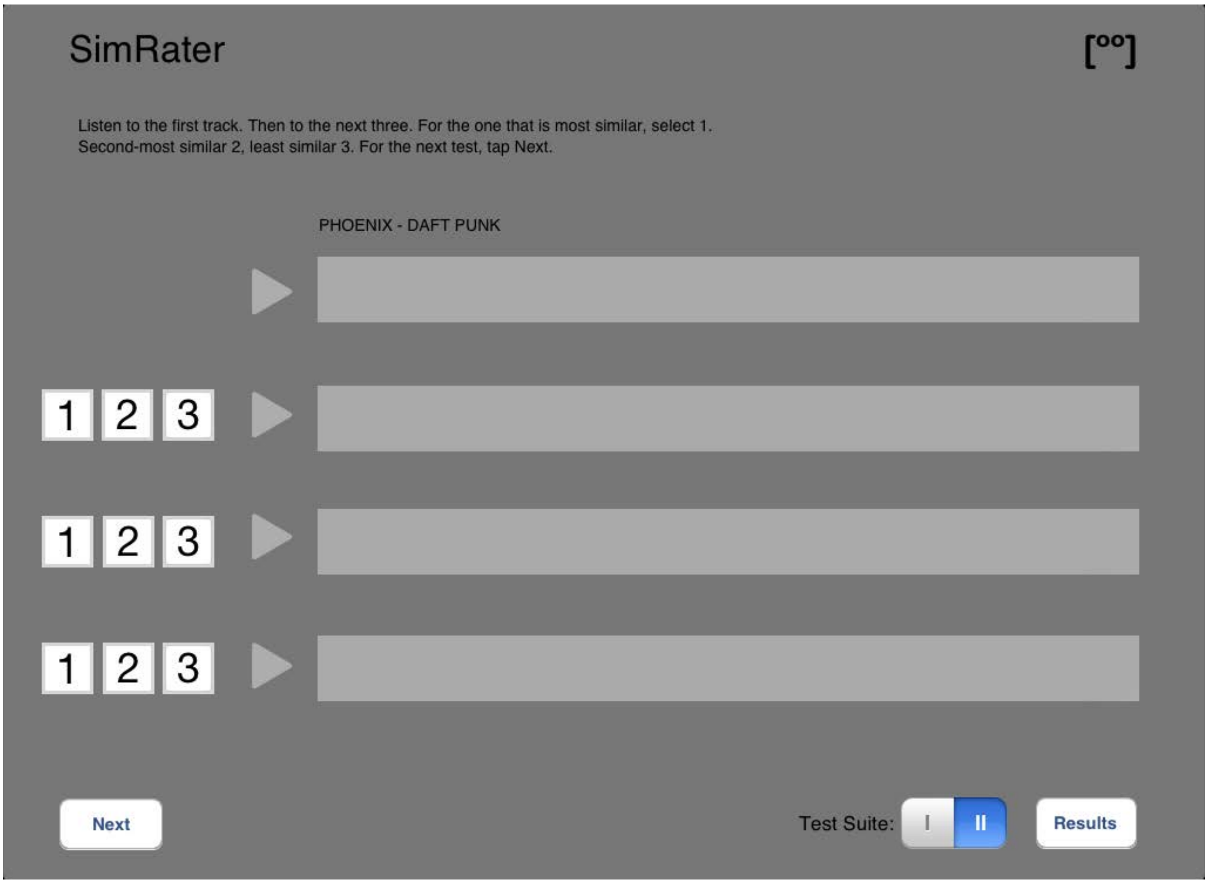
The evaluation of similarity ratings is problematic since virtually no ground truth corpora exist for these tasks, let alone for the even more specified task of timbre similarity of musical EDM segments. We have chosen here, as a preliminary attempt for evaluation, to evaluate the algorithm by testing an application based on it. We have developed an application that returns, based on an input segment of music, three other segments of music, in the order from most similar to least similar. This application can possibly be used by DJs to help them to mix music. A screenshot of the rating app is displayed in Figure 10.

To evaluate this application, we set up an experiment. We have asked subjects to, given one input segment of music, and three output segments, to order the output segments from most similar to least similar. We have averaged the ratings of the subjects and compared those to the ordering of the algorithm. We have counted the number of full ratings (the order of most to least similar) that were correctly predicted by the algorithm, as well as the number of correctly predicted most similar segments and correctly predicted least similar segments. The results can be found in table 4.

Correctness	**Full order**	**Most similar segment**	**Least similar segment**
Results	38 %	60 %	59 %
Chance level	16.7 %	33.3 %	33.3%

Table 4. Evaluation of timbre similarity algorithm. The similarity order was compared to a similarity ordering that subjects came up with.

In the experiment, 100 similarity tests have been done, by a total of five subjects. Only 38% of the full ratings were correctly predicted. A percentage of 60% correctness was obtained for scoring the most similar segment, meaning that in 22% of the cases, the least similar and 'middle' similar segment had been switched. Comparing to the level of obtaining a correct result by chance, the algorithm performed above chance level. Various reasons may exist for the fact that the performance of the algorithm does not reach high percentages. No overlap of songs over subjects was present, which means that we received only one similarity ordering per set of three output segments. Because of this last factor, validity of the created ground truth for the algorithm may be questionable. We will come back to this and other issues in the discussion section.



Listen to the first track. Then to the next three. For the one that is most similar, select 1. Second-most similar 2, least similar 3. For the next test, tap Next.

Figure 10: Screenshot of the rating iPad app

4.4. Rhythm Similarity

The MIR community has produced a lot of research on rhythm- description and similarity in the last years (Dixon, Gouyon, and Widmer, 2004; Dixon, Pampalk, and Widmer, 2003, 2004; Holzapfel, Flexer, and Widmer, 2011; Pampalk, 2006; Pampalk, Dixon, and Widmer, 2003; Pohle et al., 2009). Below we present some of the most common features in rhythm similarity.

a) Event Density

Event density is the average frequency of events, i.e. the number of onsets per second (Lartillot and Toiviainen, 2007). We compute the event density in four frequency bands, ending up with four features.

b) Fluctuation Patterns Summary

Fluctuation strength is, in principle, similar to roughness, except it quantifies subjective perception of slower (up to 20Hz) amplitude modulation of a sound (Cox, 2013). The loudness modulation has different effects on our sensation depending on the frequency. The sensation of fluctuation strength is most intense around 4Hz and gradually decreases up to a modulation frequency of 15Hz (Fastl, 1982). Fluctuation patterns describe the amplitude modulation of the loudness per frequency band (Pampalk, 2006).

We computed the fluctuation patterns using MIR Toolbox's implementation (Lartillot and Toiviainen, 2007). First the spectrogram is computed on frames of 23ms and half overlapping, then the Terhardt outer ear modeling is computed, with Bark band redistribution of the energy, and estimation of the masking effects, and finally the amplitudes are computed in dB scale. Then a FFT is computed on each Bark band, from 0 to 10 Hz, with a resolution of 0.01 Hz. The amplitude modulation coefficients are weighted based on the psychoacoustic model of the fluctuation strength (Fastl, 1982). The resulting spectrum is subsequently summed across bands, leading to a spectrum summary, showing the global repartition of rhythmic periodicities.

c) Rhythm Patterns

Pohle et al. (2009) suggested several changes to the computation of fluctuation patterns. Following one of their suggestions, we reduced the signal to the likely onsets, and created some new patterns that are robust to tempo variability, a desirable characteristic of features for computation of rhythmic similarity (Holzapfel and Stylianou, 2009).
To compute these rhythm patterns we start by dividing the audible frequency range into four bands and performing onset detection on each band. Using the tempo information estimated before, we then divide each beat into 12 bins (1 measure = 48 bins).
Then we look for onsets that repeat every measure, every two measures and every four measures. Consider for example the onset sequence that is visualized in the 4-measure pattern in figure 11, in which the vertical lines present onsets. Counting the onsets within these four measures, if we superpose the 3rd and 4th measures on the 1st and 2nd, we end up with the onset sequence that is displayed in the 2-measure pattern. If we superpose the 2nd, 3rd, and 4th measures on the 1st, we end up with the 1-measure pattern.

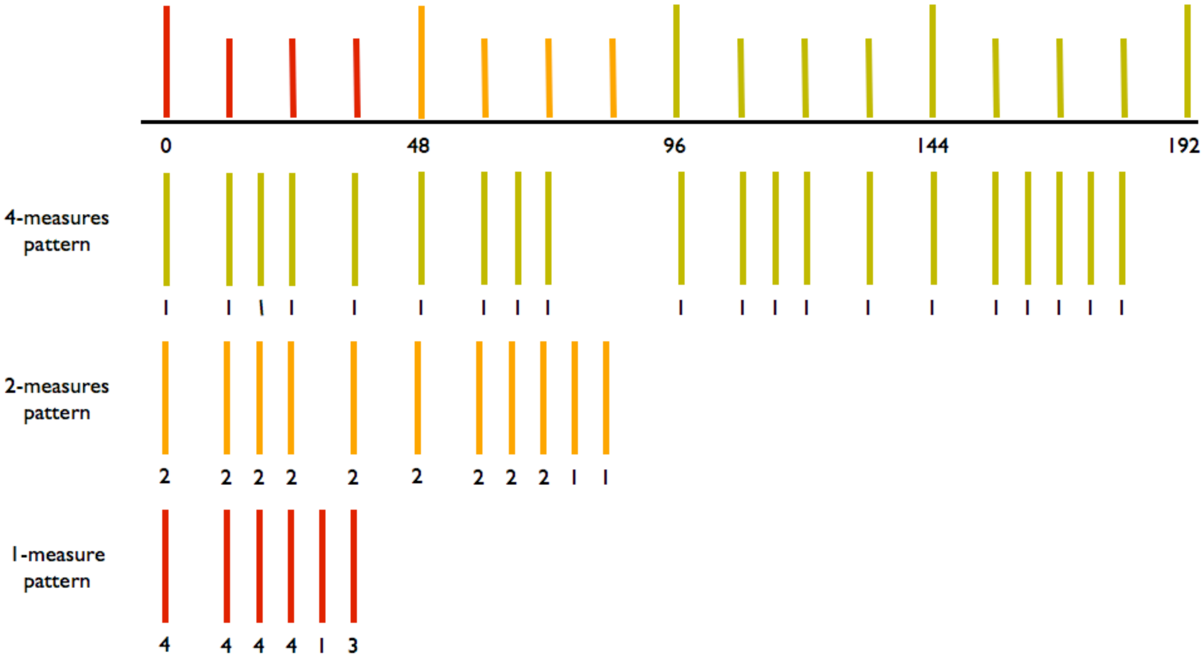


Figure 11: Rhythm patterns. Vertical lines represent onsets. On the horizontal axis are the bins (48 per measure). Numbers below the vertical lines correspond to the onsets per bin.

We repeat this process for the four frequency bands and we end up with 16 rhythm patterns for each segment. For example, the 1-measure pattern is characterized by a vector with 48 numbers from 0 to N (N = number of measures in segment) where 0 means no onsets for that bin in any measure of the segment, while N means there are onsets for that bin in all measures of the segment.

The aforementioned features (4 Event Density values, Fluctuation Pattern Summary, and 16 Rhythm Patterns) together make a feature vector that describes the rhythm of a segment.

The rhythm features have not been tested nor implemented in the final application of this project. We hope to continue the project such as to be able to work further on this.

5. Conclusions and Discussion

We have presented our model for structural segmentation and timbre similarity for electronic dance music. The segmentation algorithm was evaluated on various corpora, and performed best on an in-house dataset of EDM. Although this method was created specifically for EDM, results on the RWC Pop dataset can compete with the best performing algorithms submitted to MIREX 2012, suggesting that the structural boundaries underlying EDM follow the same principles as the boundaries in pop music. The timbre similarity algorithm has been subject to a preliminary evaluation. Since no generally accepted ground truth exists for this specific task, we have initiated a pilot experiment for this end. Although the algorithm performed above chance level, the overall performance could be improved enormously. The rhythm similarity features have not been implemented in the application and have not been tested. We plan to do this is future research. Issues related to the evaluation of segmentation and similarity will be discussed below.

In the literature, the topic of segmentation has been approached from different angles, and can be interpreted as phrasing/grouping (Bod, 2002; Lerdahl and Jackendoff, 1983), or structural segmentation (Bruderer, McKinney, and Kohlrausch, 2006; Levy and Sandler, 2008). Structural segmentation is described as to identify the key structural sections in musical audio as for example verse and chorus, and should be accessible to everybody (needing no particular musical knowledge) ("Structural Segmentation", 2012). The question however here is, whether there is indeed consensus on the concept of structural segmentation. One issue that we came across, for example, is the question whether it is necessary for a segment boundary to coincide with a downbeat. We found this to be the case for EDM, and in this case the "preference" for perceiving a new segment starting on a downbeat overruled the concept of timbral change that was underlying our algorithm, hence the introduction of the musically informed rules. One can wonder whether this is the case for other genres as well.

An issue related to this is how far phrase-segmentation and structural segmentation merge. If a phrase, starting with an upbeat, introduces the start of new structural segment, does the structural segment start with the start of the phrase (on the upbeat) or does it start on the downbeat following the upbeat?

The evaluation process of segmentation algorithms is important to consider as well. Several studies use only a ± 3s tolerance window for evaluation. The question is whether this window is not too large to be able to assess algorithms in a detailed way. If the large window is used to cover up misalignments like ones caused by issues that we outlined above (e.g. boundaries on upbeats or downbeats), then these are the issues that we should consult instead of hiding them with large tolerance windows. Problems like these have been discussed before (Rocha et al., 2012) and we feel it is important to continue this discussion.

With respect to similarity in music, we argued that similarity should always have a context. Therefore we have broken the total concept of similarity into sub-similarities of which timbre similarity is one. The most common method is to evaluate a model against a manually annotated ground truth. Here the question is whether it makes sense to ask subjects to rate timbre similarity. Is it possible to assess only timbre similarity, leaving all other sub-similarities, such as melodic-similarity, rhythmic-similarity, etc. out? Although we argued that timbre is the primary compositional parameter for EDM, analysis of the results of our experiment suggests that this is probably a hard thing to do. As Aucouturier and Pachet (2002) point out, people's similarity judgments are simultaneously influenced by other factors.

Besides the question whether it is possible to assess timbre similarity alone, another question is whether it is possible to perceive a total concept of timbre. If we think of a piece of music consisting of, for example, simultaneous saxophone-, drum- and piano-parts, we most probably perceive three different streams (cf. Cambouropoulos, 2008). So if we perceive three different timbres, is it possible as well to perceive one overarching timbre that covers the three timbres? And if not, how do we compare the timbre of a piece of music to another piece of music? It is known that timbre is a multidimensional concept (McAdams et al., 1995) and the question is whether it should be assessed in this way as well.

With respect to the evaluation of the timbre similarity algorithm in this paper, there are a number of additional issues. In this report, only a preliminary evaluation of the algorithm is presented, and we plan to follow up on this in a future study. A ground truth was created by five people who rated a total of 100 sets of segments, from most to least similar with respect to a query-segment. No overlap existed between the music that the five people rated, which means that each set of segments was rated by only one person. Therefore, the confidence of every rating is rather low.

Another issue is that of the temporal evolution of the timbre over a segment. The motivation for segmenting the music was that timbres could change a lot over the course of a whole song. However, it is possible as well for the timbre to change within a segment. Imagine the following typical scenario: a segment is composed of 8 measures, of which the last 2 have no bass or kick drum; computing the average low-frequency energy of the segment will give us a value that is not representing well the *feeling* of the segment. The present algorithm is computing an average timbre over a segment, but we plan to take into account the temporal evolution of the timbre over a segment in future research.

The algorithm in its present form computes a similarity rating based on different timbral features with equal weighting. One might expect that some features may be more important than others and that the optimal weighting scheme is different from the one we have used here. We plan to do a full evaluation of this similarity measure by creating a bigger ground truth corpus in the near future. This evaluation will also include statistical tests involving other features, optimization of feature weighting, and comparisons between MFCC-only approaches (e.g. Terasawa et al., 2005) and ours.

Acknowledgments

This research has been funded by the Center for Digital Humanities and the Netherlands foundation of Scientific Research (NWO), grant no. 639.021.126.

6. References

Allamanche, E., Herre, J., Hellmuth, O., Kastner, T., & Ertel, C. (2003). A multiple feature model for musical similarity retrieval. *Proceedings of the 4th International Conference on Music Information Retrieval*. Baltimore, USA.

Alluri, V., & Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, 27(3), 223–242.

Aspillaga, F. X., Cobb, J., & Chuan, C. H. (2011). Mixme: A recommendation system for DJs. *Late-break Session of the 12th International Society for Music Information Retrieval Conference*. Miami, USA.

Aucouturier, J. J., & Pachet, F. (2002). Music similarity measures: What's the use. In *Proceedings of the 3rd International Conference on Music Information Retrieval*. Paris, France.

Aucouturier, J. J. & Pachet, F. (2004a). Improving timbre similarity: How high is the sky? *Journal of negative results in speech and audio sciences*, 1(1): 1–13.

Aucouturier, J. J., & Pachet, F. (2004b). Tools and architecture for the evaluation of similarity measures: case study of timbre similarity. *Proceedings of the 5th International Conference on Music Information Retrieval*. Barcelona, Spain.

Aucouturier, J. J., & Sandler, M. (2001). Segmentation of Musical Signals Using Hidden Markov Models. *Proceedings of the 110th Convention of the Audio Engineering Society*. Amsterdam, The Netherlands.

Aucouturier, J. J., Pachet, F., & Sandler, M. (2005). "The Way It Sounds": Timbre Models for Analysis and Retrieval of Music Signals. *IEEE Transactions on Multimedia*, 7(6): 1028-1035.

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions. *Science*, 338(6111): 1225–1229.

Berenzweig, A., Ellis, D. P., & Lawrence, S. (2003). Anchor space for classification and similarity measurement of music. *Proceedings of the IEEE International Conference on Multimedia and Expo*. Baltimore, USA.

Berenzweig, A., Logan, B., Ellis, D. P. W., & Whitman, B. (2004). A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures. *Computer Music Journal*. 28(2): 63-76.

Bimbot, F., Deruty, E., Sargent, G., & Vincent, E. (2011). Methodology and resources for the structural segmentation of music pieces into autonomous and comparable blocks. *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, USA.

Bimbot, F., Le Blouch, O., Sargent, G., & Vincent, E. (2010). Decomposition into autonomous and comparable blocks: a structural description of music pieces. *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Utrecht, The Netherlands.

Bod, R. (2002). Memory-based models of melodic analysis: Challenging the Gestalt principles. *Journal of New Music Research*, 31(1), 27-36.

Bogdanov, D., Serrà, J., Wack, N., Herrera, P., & Serra, X. (2011). Unifying Low-Level and High-Level Music Similarity Measures. *IEEE Transactions on Multimedia*, 13(4): 687-701.

Bogert, B. P., Healy, M. J. R., & Tukey, J. W. (1963). The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In M. Rosenblatt (Ed.), *Time Series Analysis* (pp. 209-243).

Bruderer, M. J., McKinney, M., & Kohlrausch, A. (2006). Structural boundary perception in popular music. *Proceedings of the 7th International Conference on Music Information Retrieval* (pp. 198-201). Victoria, Canada.

Burgoyne, J., & McAdams, S. (2008). A meta-analysis of timbre perception using nonlinear extensions to CLASCAL. *Proceedings of the 5th International Symposium on Computer Music Modeling and Retrieval*. Copenhagen, Denmark.

Butler, M. J. (2006). *Unlocking the Groove: Rhythm, Meter, and Musical Design in Electronic Dance Music*. Bloomington, USA: Indiana University Press.

Cambouropoulos, E. (2008). Voice and stream: Perceptual and computational modeling of voice separation. *Music Perception*, 26(1): 75-94.

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-Based Music Information Retrieval: Current Directions and Future Challenges. *Proceedings of the IEEE*, 96(4): pp. 668-696.

Chen, R., & Li, M. (2011) Music structural segmentation by combining harmonic and timbral information. *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, USA.

Chew, E., Volk, A., & Lee, C. Y. (2005). Dance Music Classification Using Inner Metric Analysis: a computational approach and case study using 101 Latin American Dances and National Anthems. In B. Golden, S. Raghavan, & E. Wasil (Eds.), *The Next Wave in Computing, Optimization, and Decision Technologies* (pp. 355-370). New York, USA: Springer.

Cilibrasi, R., Vitanyi, P., & de Wolf, R. (2004). Algorithmic clustering of music based on string compression. *Computer Music Journal*, 28(4): 49-67.

Cox, T. (2013, March 7). Fluctuation Strength. *University of Salford.* Retrieved from http://www.acoustics.salford.ac.uk/res/cox/sound_quality/?content=roughness

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4): 357–366.

De Leon, F., & Martinez, K. (2012). Enhancing timbre model using MFCC and its time derivatives for music similarity estimation. *Proceedings of the 20th European Signal Processing Conference*. Bucharest, Romania.

De Nooijer, J., Wiering, F., Volk, A., & Tabachneck-Schijf, H. J. (2008). An experimental comparison of human and automatic music segmentation. *Proceedings of the 10th International Conference on Music Perception and Cognition*. Sapporo, Japan.

Deliège, I. (2001). Similarity Perception ↔ Categorization ↔ Cue Abstraction. *Music Perception*, 18(3), 233–243.

Dixon, S., Gouyon, F., & Widmer, G. (2004). Towards characterisation of music via rhythmic patterns. In *Proceedings of the 5th International Conference on Music Information Retrieval.* Barcelona, Spain.

Dixon, S., Pampalk, E., & Widmer, G. (2003). Classification of dance music by periodicity patterns. *Proceedings of the 4th International Conference on Music Information Retrieval* (pp. 159-165). Baltimore, USA.

Dixon, S., Pampalk, E., & Widmer, G. (2004). Evaluating rhythmic descriptors for musical genre classification. *Proceedings of the 25th International Audio Engineering Society Conference: Metadata for Audio*. London, UK.

Ehmann, A., Bay, M., Downie, J. S., Fujinaga, I., & De Roure, D. (2011). Music Structure Segmentation Algorithm Evaluation: Expanding on MIREX 2010 Analyses and Datasets. *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, USA.

Ellis, D. P., Whitman, B., Berenzweig, A., & Lawrence, S. (2002). The quest for ground truth in musical artist similarity. *Proceedings of the 3rd International Conference on Music Information Retrieval.* Paris, France.

Fales, C. (2002). The paradox of timbre. *Ethnomusicology*, 46(1), 56–95.

Fales, C. (2004). "Heaviness" in the Perception of Heavy Metal Guitar Timbres The Match of Perceptual and Acoustic Features over Time**.** In P. Greene & T. Porcello (Eds.), *Wired For Sound: Engineering And Technologies In Sonic Cultures* (pp. 181-197). Middletown, USA: Wesleyan University Press.

Fastl, H. (1982). Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise. *Hearing Research*, 8(1): 59-69.

Ferguson, J. P. (2012, May 31). EDM is taking over the Chicago festival season. *Time Out Chicago*. Retrieved from http://timeoutchicago.com.

Fikentscher, K. (2000). *"You Better Work!": Underground Dance Music in New York*. Middletown, USA: Wesleyan University Press.

Flexer, A., Schnitzer, D., & Schlüter, J. (2012). A MIREX meta-analysis of hubness in audio music similarity. *Proceedings of the 13th international society for music information retrieval conference.* Porto, Portugal.

Flexer, A., Schnitzer, D., Gasser, M., & Widmer, G. (2008). Playlist generation using start and end songs. *Proceedings of the 9th International Conference on Music Information Retrieval*. Philadelphia, USA.

Foote, J. T. (1997). Content-based retrieval of music and audio. *Proceedings of Multimedia Storage and Archiving Systems II* (SPIE 3229).

Foote, J. T. (2000). Automatic audio segmentation using a measure of audio novelty. *Proceedings of the IEEE International Conference on Multimedia and Expo*. New York, USA.

Foote, J. T., & Cooper, M. L. (2003). Media segmentation using self-similarity decomposition. *Proceedings of Storing and Retrieval for Media Databases* (SPIE 5021).

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K.J. Holyoak & R.G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 13-36). Cambridge, UK: Cambridge University Press.

Goodman, N. (1972). Seven strictures on similarity. In *Problems and Projects* (pp. 437-447). Indianapolis, USA: Bobbs-Merrill.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC Music Database: Popular, Classical, and Jazz Music Databases. *Proceedings of the 3rd International Conference on Music Information Retrieval*. Paris, France.

Greenburg, Z. (2012, August 2). The world's highest-Paid DJs 2012. *Forbes*. Retrieved from http://www.forbes.com.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61: 1270-1277.

Grosche, P., Müller, M., & Serrà, J. (2012). Audio Content-Based Music Retrieval. *Multimodal Music Processing*, 3: 157–174.

Hellman, R. P. (1972). Asymmetry of masking between noise and tone. *Attention, Perception, & Psychophysics*, 11(3): 241–246.

Helmholtz, H. (1954). *On the Sensations of Tone*. New York, USA: Dover.

Herre, J., Allamanche, E., & Ertel, C. (2003). How similar do songs sound? Towards modeling human perception of musical similarity. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

Herre, J., Allamanche, E., & Hellmuth, O. (2001) Robust Matching Of Audio Signals Using Spectral Flatness Features. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

Hesmondhalgh, D. (1998). The British dance music industry: a case study of independent cultural production. *British Journal of Sociology*, 234–251.

Hillewaere, R., Manderick, B., & Conklin, D. (2010). String quartet classification with monophonic models. *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Utrecht, The Netherlands.

Holzapfel, A., & Stylianou, Y. (2009). A scale transform based method for rhythmic similarity of music. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 317-320). Taipei, Taiwan.

Holzapfel, A., Flexer, A., & Widmer, G. (2011). Improving tempo-sensitive and tempo-robust descriptors for rhythmic similarity. *Proceedings of the 8th Sound and Music Computing Conference.* Padova, Italy.

Honingh, A. & Bod, R. (2011). Clustering and classification of music using interval categories. *Proceedings of the 3rd International Conference on Mathematics and Computation in Music*. Paris, France.

Jayant, N. S. & Noll, P. (1984) *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, USA: Prentice-Hall.

Jensen, J. H., Christensen, M. G., Ellis, D. P. W., & Jensen, S. H. (2009). Quantitative Analysis of a Common Audio Similarity Measure. *IEEE Transactions on Audio, Speech, and Language Processing,* 17(4): 693–703.

Kaminskas, M., & Ricci, F. (2012). Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, 6(2-3): 89–119.

Lartillot, O., & Toiviainen, P. (2007). A Matlab Toolbox for Musical Feature Extraction from Audio. *Proceedings of the 10th International Conference on Digital Audio Effects.* Bordeaux, France.

Lerdahl, F., Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, USA: MIT Press.

Levy, M., & Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing,* 16(2): 318-326.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. *Proceedings of the 1st International Symposium on Music Information Retrieval*. Plymouth, USA.

Logan, B., & Salomon, A. (2001). A music similarity function based on signal analysis. *Proceedings of the IEEE International Conference on Multimedia and Expo*. Tokyo, Japan.

Logan, B., Ellis, D. P., & Berenzweig, A. (2003). Toward evaluation techniques for music similarity. *Proceedings of the Workshop on the Evaluation of Music Information Retrieval Systems.* Toronto, Canada.

Marui, A., & Martens, W. L. (2005). Timbre of nonlinear distortion effects: Perceptual attributes beyond sharpness. *Proceedings of the 2nd Conference of Interdisciplinary Musicology*. Montreal, Canada.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3), 177–192.

McFee, B. (2012). *More like this: machine learning approaches to music similarity.* Doctoral Dissertation. University of California, San Diego, USA.

McLeod, K. (2001). Genres, Subgenres, Sub-Subgenres and More: Musical and Social Differentiation Within Electronic/Dance Music Communities. *Journal of Popular Music Studies,* 13: 59-75.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review*, 100(2): 254.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, 92(3), 289.

Musil, J. J., Elnusairi, B., & Müllensiefen, D. (2012). Perceptual dimensions of short audio clips and corresponding timbre features. *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval.* London, UK.

Navarro, D. J., & Lee, M. D. (2003). Combining Dimensions and Features in Similarity Based Representations In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 59–66). Cambridge, USA: MIT Press.

Oppenheim, A. V., & Schafer, R. W. (2004). From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, 21(5): 95–106.

Pampalk, E. (2004). A Matlab toolbox to compute music similarity from audio. In *Proceedings of the 5th International Conference on Music Information Retrieval*. Barcelona, Spain.

Pampalk, E. (2006). *Computational models of music similarity and their application in music information retrieval*. Doctoral dissertation. Vienna University of Technology, Austria.

Pampalk, E., Dixon, S., & Widmer, G. (2003). On the evaluation of perceptual similarity measures for music. *Proceedings of the 6th International Conference on Digital Audio Effects*. London, UK.

Pampalk, E., Flexer, A., & Widmer, G. (2005). Improvements of audio-based music similarity and genre classification. *Proceedings of the 6th International Conference on Music Information Retrieval*. London, UK.

Patil, K., Pressnitzer, D., Shamma, S., & Elhilali, M. (2012). Music in Our Ears: The Biological Bases of Musical Timbre Perception. *PLoS Computational Biology*, 8(11), e1002759.

Paulus, J., & Klapuri, A. (2009). Music Structure Analysis Using a Probabilistic Fitness Measure and a Greedy Search Algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6): 1159–1170.

Paulus, J., Müller, M., & Klapuri, A. (2010). State of the art report: Audio-based music structure analysis. *Proceedings of the 11th International Society for Music Information Retrieval Conference.* Utrecht, The Netherlands.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical Report. IRCAM, France.

Peeters, G. (2007). Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. *Proceedings of the 8th International Conference on Music Information Retrieval*. Vienna, Austria.

Peeters, G., & Deruty, E. (2009). Is music structure annotation multidimensional? A proposal for robust local music annotation. *Proceedings of the 3rd Workshop on Learning the Semantics of Audio Signals.* Graz, Austria.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, 130(5): 2902-2916.

Peiszer, E., Lidy, T., & Rauber, A. (2008). Automatic audio segmentation: Segment boundary and structure detection in popular music. *Proceedings of the 2nd International Workshop on Learning Semantics of Audio Signals*. Paris, France.

Perez-Sancho, C., Rizo, D., & Inesta, J. (2009). Genre classification using chords and stochastic language models. *Connection Science*, 21(2-3): 145-159.

Plomp, R., & Levelt, W. J. (1965). Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America*, 38(4): 548–560.

Pohle, T., Schnitzer, D., Schedl, M., Knees, P., & Widmer, G. (2009). On rhythm and general music similarity. *Proceedings of the 10th International Conference on Music Information Retrieval.* Kobe, Japan.

Pollastri, E. & Simoncelli, G. (2001). Classification of melodies by composer with hidden Markov models. *Proceedings of the 1st International Conference on Web Delivering of Music*.

Pope, S. T. (2009). Improving Music Information Retrieval using Segmentation-related Statistics. Presented at *The Future of Interactive Media: Workshop on Media Arts, Science, and Technology*. Santa Barbara, USA.

Prünster, G., Fellner, M., Graf, F., & Mathelitsch, L. (2004). An empirical study on the sensation of roughness. *Proceedings of the 1st Conference on Interdisciplinary Musicology.* Graz, Austria.

Rocha, B., Smith, J. B. L., Peeters, G., Ross, J. C., Nieto, O., & van Balen, J. (2012). Late-break session on music structure analysis. *Late-break Session of the 13th International Society for Music Information Retrieval Conference.* Porto, Portugal.

Ruwet, N., & Everist, M. (1987). Methods of analysis in musicology. *Music Analysis*, 6(1/2): 3-36.

Ryce, A. (2012, September 11). RA Roundtable: EDM in America. *Resident Advisor*. Retrieved from http://www.residentadvisor.net.

Sanden, C., Befus, C. R., & Zhang, J. Z. (2012). A Perceptual Study on Music Segmentation and Genre Classification. *Journal of New Music Research*, 41(3): 277–293.

Schnitzer, D., Flexer, A., Schedl, M., & Widmer, G. (2011) Using mutual proximity to improve content-based audio similarity. *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, USA.

Schwarz, D., & Hackbarth, B. (2012). Navigating variation: composing for audio mosaicing. *Proceedings of the International Computer Music Conference*. Ljubljana, Slovenia.

Serrà, J., Muller, M., Grosche, P., & Arcos, J. L. (2012). Unsupervised Detection of Music Boundaries by Time Series Structure Features. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 1613-1619. Toronto, Canada.

Sethares, W. A. (1998). *Tuning, Timbre, Spectrum, Scale*. New York, USA: Springer.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers,* 37(1): pp. 10-21.

Shepard, R. N. (1962 a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2): 125–140.

Shepard, R. N. (1962 b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27(3): 219–246.

Smith, J. B. L., Burgoyne, J. A., De Roure, D., Fujinaga, I., & Downie, J. S. (2011). Design and creation of a large-scale database of structural annotations. *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, USA.

Smith, J. O. (2011). Spectrum Analysis Windows. *Spectral Audio Signal Processing*. Retrieved from http://ccrma.stanford.edu/~jos/sasp/.

Spiro, N. (2007). *What contributes to the perception of musical phrases in western classical music?* Doctoral Dissertation. University of Amsterdam, The Netherlands.

Structural Segmentation (2012). Retrieved from http://www.music-ir.org/mirex/wiki/2012:Structural_Segmentation

Styles of house music. (2013, March 7). *Wikipedia, The Free Encyclopedia*. Retrieved from http://en.wikipedia.org.

Terasawa, H., Slaney, M., and Berger, J. (2005). Perceptual distance in timbre space. *Proceedings of the International Conference on Auditory Display (ICAD05)*, pp. 1-8.

Tillmann, B., & McAdams, S. (2004). Implicit Learning of Musical Timbre Sequences: Statistical Regularities Confronted With Acoustical (Dis)Similarities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(5), 1131–1142.

Toiviainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huotilainen, M., & Näätänen, R. (1998). Timbre similarity: Convergence of neural, behavioral, and computational approaches. *Music Perception*, 223–241.

Tversky, A. (1977). Features of Similarity. *Psychological Review*, 84(4): 327-352.

Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and categorization*, 1: 79–98.

Tzanetakis, G., & Cook, P. (2000). Audio information retrieval (AIR) tools. *Proceedings of the 1st International Symposium on Music Information Retrieval.* Plymouth, USA.

Vassilakis P. N. and Kendall, R. A. (2010). Psychoacoustic and cognitive aspects of auditory roughness: definitions, models, and applications. *Proceedings of Human Vision and Electronic Imaging XV* (SPIE 7527).

Vassilakis, P. N. (2001). *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance* (Eq. 6.23). Doctoral Dissertation. University of California, Los Angeles, USA.

Volk, A., de Haas, B., & van Kranenburg, P. (2012). Towards Modelling Variation in Music as Foundation for Similarity. *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music*. Thessaloniki, Greece.

Weiss, R. J., & Bello, J. P. (2011). Unsupervised Discovery of Temporal Structure in Music. *IEEE Journal of Selected Topics in Signal Processing*, 5(6): 1240–1251.

West, K. (2008). *Novel techniques for audio music classification and search*. Doctoral Dissertation. University of East Anglia, UK.

West, K., & Lamere, P. (2007). A model-based approach to constructing music similarity functions. *EURASIP Journal on Applied Signal Processing*, 1: 149–149.

Wiggins, G. (2007). Models of musical similarity. *Musicae Scientiae*, Discussion Forum 4a, pp. 315-338.

Yenigun, S. (2012, October 29). Dance music looks beyond EDM and hopes the crowd will follow. *NPR*. Retrieved from http://www.npr.org.

Yeston, M. (1976). *The Stratification of Musical Rhythm* (p. 41). New Haven, USA: Yale University Press.

Zacharakis, A., Pastiadis, K., Papadelis, G., & Reiss, J. D. (2011). An investigation of musical timbre: uncovering salient semantic descriptors and perceptual dimensions. *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Miami, USA.