ABSTRACT

In this paper we argue that the notion of music similarity should be expanded into sub-similarities, meaning that similarity of music has to be judged with respect to a certain context, such as melody, harmony, rhythm or timbre. We start by focusing on timbre similarity, restricted to the domain of Electronic Dance Music (EDM). We will assess the similarity of segments of music, thus we start by studying segmentation before we come to the topic of similarity. The segmentation algorithm performs well on an EDM dataset as well as on a standard MIREX dataset. Initial listening tests of the similarity model give promising results but will have to be further evaluated in future research.

1. INTRODUCTION

Similarity in music is a fascinating but 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.

In the academic field of Music Information Retrieval, various systems have been developed that classify music according to a certain type of similarity [13,14,25,35,37]. 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 or listener ratings 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 apps use metadata (keywords tagged by the user which can include information about artist, title, genre and more [29]), or collaborative filtering (relevance of a song to a user is predicted based on similar users’ ratings [25]), meaning that the music itself (the audio file) is not studied. 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. The presented research in this paper focuses on content-based music retrieval, in which the audio is studied. In this way, we can use all music that we have, 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, it is known that similarity depends on context [11]. By all means, we can imagine that a piece of music could be rhythmically similar to another piece, without being similar in melody or harmony (e.g. salsa music has a typical rhythm, similar for most salsa music, but different salsa tracks vary in melody/harmony/...). Therefore, it is useful to expand the notion of similarity into sub-similarities, meaning that similarity of music has to be judged with respect to a certain context, such as melody, harmony, rhythm or timbre. The research described in this paper is part of a larger project in which several sub-similarities are studied. This paper focuses on timbre similarity.

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 have focused on the similarity of segments of music, and thus we start by studying segmentation before we come to the topic of similarity.

Since the topic of music similarity, even when restricted to just timbre 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. After a brief introduction into EDM, this paper will report on our study on segmentation and timbre similarity.

The contributions of this paper are two-fold: (1) we present an algorithm for the detection of structural boundaries in EDM, of which the main innovations are the addition of a first downbeat detection and the implementation of musically informed rules; (2) an algorithm to perform polyphonic timbre similarity is presented, of which the main novelty is the modification of the concept of roughness.
2. ELECTRONIC DANCE MUSIC (EDM)

Electronic Dance Music (EDM) is a label that defines a metagene encompassing a heterogeneous group of musics made with computers and electronic instruments [27]. 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 [9].

Until recently, EDM was (with some sporadic exceptions) an underground culture, i.e., cultivated outside the view of the general public eye [18], but it has risen to the mainstream charts of the music industry [20]. 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 soundtrack of choice for a new generation” [17].

Almost all EDM share certain musical characteristics: (1) a steady tempo, mostly in the range of 120-150 BPM dependent of genre; (2) a repeating bass drum pattern [9].

Timbre, often referred to as texture, stands out as a primary compositional parameter in EDM. It is seen as the criterion by which patterns may be differentiated most easily [46]. Most of the timbral changes that occur in EDM involve an element either entering or leaving the mix. In Butler [9], DJs Shiva and Stanley described a prototypical structure of EDM tracks. They based their descriptions mainly on timbral changes. As the DJs Butler [9] 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 [39]. 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 [12]. Tzanetakis and Cook [43] stress the importance of segmentation in Music Information Retrieval (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 [12].

As pointed out before, timbral changes are essential for EDM producers when considering structural changes. Au-courtier and Sandler [2] argue that, to segment a song into its relevant sections, one should discard any pitch and harmonic information and focus on timbre alone.

To find structural segments in EDM we will (1) extract timbral 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 [28]. 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. We will explain all steps of the segmentation algorithm below. The MIR Toolbox [28] was used to perform most of the steps.

3.1 Detection of first bass drum downbeat

Many EDM tracks begin with beatless intros and culminate in turning the beat around, a phenomenon that occurs when people perceive a certain metrical structure which is violated later (usually by introducing a beat on the perceived off-beat) [9]. For this reason, the entrance of the bass drum in an EDM track often results in a decisive metrical representation [9]. 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 and 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. See figure 1 for a visual explanation.

3.2 Tempo estimation and confidence measure

Tempo estimation is performed in order to detect the duration of a beat. This is important because all features (for both the segmentation and the similarity tasks) are extracted on beat-related frame lengths.

Looking at local correlation between samples we can evaluate periodicities in a signal. 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 [28]. 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 selected as the tempo of the track. A binary confidence measure telling us about the likelihood 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. If there are several peaks with no har-
monic relation between the spacing of the peaks, the estimated confidence value is 0. This measure will be used later, on the level of fine-tuning the segment boundaries (section 3.4).

### 3.3 Novelty detection

After computing the tempo score in beats per minute (BPM) and building a vector with all the beat positions, we compute the magnitude spectrum of each frame of the signal. The frames are beat-aligned with 87.5% overlap so that we decompose the energy along frequencies for each beat of the track.

We perform a cepstrum analysis in order to find periodic sequences in the signal. This is motivated by the fact that timbre should be the most important characteristic for segmentation [2, 9], and by analysis on both MFCC and cepstrum-based segmentation.

Following Foote [19], we then compute the cosine distance between each possible pair of frames from the cepstrum data to get a self-similarity matrix (figure 2). Convoluting along the main diagonal of the similarity matrix results in a novelty curve that indicates the temporal locations of significant timbral changes by its peaks. These locations present the segment boundaries that we searched for.

### 3.4 Musically informed rules

Although our algorithm located the segment boundaries based on timbral changes, we are not done yet. The nature of EDM requires us to fine-tune the segment boundary locations. Butler [9] 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 multi-measure group, in order to raise expectation for a segment boundary for the listener. 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 usually follow these sounds and start on a downbeat.

To overcome this displacement, we propose a set of heuristic rules to align the obtained novelty peaks with the most probable structural boundaries - 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), 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 3 shows the effect of the rules on a hypothetical track.

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.

3.5 Evaluation of segmentation

The segmentation algorithm was evaluated using several datasets: (1) EDMs, an in-house EDM dataset specially created for this project, consisting of 35 songs - annotated by the authors - from 19 artists; (2) RWC Pop [21], annotated by two groups of researchers RWO corresponds to the annotations of the dataset creators and RWQ corresponds to the annotations made in the Quaero project [4]; (3) Eurovision dataset [5]. Found segment boundaries are considered correct if they are within ±0.5 seconds (precise) or ±3 seconds (relaxed) 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.

The results can be found in table 1. Of the EDMs dataset, we show the results both with and without the musically informed rules. On the other datasets, the rules did not make a significant difference and we show only the results where the musically informed rules have been applied. The algorithm performs well on the EDMs 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 [32]. This suggests that structural changes in pop music might have the same periodicity as in EDM. This method does not reach high performance 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. MUSIC SIMILARITY: TIMBRE

Cambouropoulos [11] explains that the concept of similarity always depends on context, such that we can only speak of similar music with respect to a certain context such as timbre, melody, rhythm etc. To make a start with defining musical similarity in this way, and since timbre is seen as the criterion by which patterns in EDM may be differentiated most easily [46], this paper focuses on timbre similarity.

Studies in timbre perception indicate 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 [16]. These factors have been identified to include, among others, spectral flux, spectral centroid, and attack time [10, 26, 34]. These studies focused on monophonic timbres. However, here we want to describe polyphonic textures, which the aforementioned features cannot represent.

For our purposes, we have empirically made a selection of a small number of features to describe 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.

4.1 Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) [15] are used 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 [28].

The number of MFCCs that well represent a spectral envelope is subject to discussion. The low order MFCCs account for the slowly changing spectral envelope, while the higher order ones describe the fast variations of the spectrum [1]. 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 [3]. The same authors reported an ideal value of 20 coefficients, which we selected.

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.

### 4.2 Spectral Flatness

Spectral flatness, also referred to as tonality coefficient, measures the sinusoidality of a spectrum [33]. It indicates whether the distribution of the spectrum is smooth or spiky, and provides a way to quantify how tone-like a sound is, as opposed to being noise-like. The spectral flatness is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum, i.e.:

$$\sqrt[\frac{N}{\prod_{n=0}^{N-1} x(n)}} \frac{1}{N} \sum_{n=0}^{N-1} x(n),$$  \hspace{1cm} (1)

where $x(n)$ represents the magnitude of bin number $n$.

For the computation of the spectral flatness, we split the spectrum into four bands. 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.

### 4.3 Dirtiness

Helmholtz [22] 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 [36]. Roughness can be considered as 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. Optimizing the weighting is planned for future research as well, but is however highly dependent on confident ground truth.

### 5. DISCUSSION AND CONCLUSIONS

We have presented our model for structural segmentation and timbre similarity for electronic dance music. The segmentation algorithm included a set of musically informed rules to account for the fact that segment boundaries in EDM are usually on the beat. The 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.

In the literature, the topic of segmentation has been approached from different angles, and can be interpreted as phrasing/grouping [6, 30], or structural segmentation [8, 31]. 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) [32]. 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 im-
important to consider as well. Several studies use only a ± 3 second tolerance window for evaluation. We would like to argue that this window is 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 [38] and we feel it is important to continue this discussion.

Besides the segmentation algorithm we have presented our model for timbre similarity in EDM. A feature vector has been created to describe a particular timbre, with the most novel feature being ‘dirtiness’, which accounts for the rough sound that is characteristic for some types of EDM. The selection of features for this feature vector was based on empirical tests on a reduced dataset, for which only the features described in section 4 seemed to reveal any particular relevance. Initial listening test gave promising results, but since no groundtruth dataset exist, a formal evaluation has not been done. We plan to do a full evaluation of this similarity measure by creating our own groundtruth corpus in the near future. This evaluation will also include statistical tests involving other features and comparisons between MFCC-only approaches (e.g. [42]) and ours.

Acknowledgments

The authors wish to thank three anonymous reviewers. This research has been funded by the Center for Digital Humanities (http://cdh.uva.nl/) and the Netherlands foundation of Scientific Research (NWO), grant no. 639.021.126.

6. REFERENCES


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


