Article

MUSICÆ SCIENTIÆ

Musicae Scientiae
1–24

DOI: 10.1177/1029864916655596
msx.sagepub.com
SAGE

A model for rhythm and timbre similarity in electronic dance music

Maria Panteli
University of Amsterdam, the Netherlands; Queen Mary University of London, UK

Bruno Rocha
University of Amsterdam, the Netherlands; University of Coimbra, Portugal

Niels Bogaards
Elephantcandy, the Netherlands

Aline Honingh
University of Amsterdam, the Netherlands

Abstract

Music similarity is a multidimensional concept to which so-called "sub-similarities", such as timbre and rhythm similarity, contribute. In this study, two models are presented: one for timbre similarity, and one for rhythm similarity. The musical domain for which the models were established is Electronic Dance Music (EDM). The models extract feature values from segments of audio and calculate a distance between two segments based on their feature vectors. The models are evaluated on perceptual data using linear regression. The accuracy of the rhythm similarity model reaches an empirically established upper bound to model performance. The accuracy of the timbre model is moderate, possibly due to insufficient data. From the selection of features and their weights resulting from the regression analysis, periodicity of rhythmic elements turned out to be the most important feature group for rhythm similarity in EDM.

Keywords

content-based, electronic dance music, music similarity, rhythm, timbre

With the growing amount of musical data available on the internet and in private collections, the need for organizing and retrieving music has increased. Models of similar music retrieval, music identification, and (genre) classification come into play here, to provide the end users with ways to address specific music requests or to explore a data collection in a systematic way.

Corresponding author:
Maria Panteli, Queen Mary University of London, Mile End Road, London, E1 4NS, UK.
Email: m.x.panteli@gmail.com

Models of music similarity can be based on a) audio content, such as when audio features are extracted and compared (Grosche, Müller, & Serrà, 2012), b) descriptive metadata, which includes information such as song title, album, artist, but also mood, genre, style, etc. (Lamere, 2008), and c) playlist information from internet users (Schedl & Knees, 2009). Each of these three methods has shown to be useful for modeling specific aspects of music similarity. In this research we are interested in a detailed level of similarity, and we want to be able to measure similarity between fragments of the same song. Therefore, we concentrate on content-based models from this point onwards.

Music similarity is by now one of the most popular topics in the field of Music Information Retrieval (MIR). Most models that use a content-based retrieval approach extract global audio features describing the complete song (Novello, van de Par, McKinney, & Kohlrausch, 2013; Pampalk, 2004; Schnitzer, Flexer, & Widmer, 2012). Since certain features might be more important for music similarity than others, various studies have tried to optimize the weights of the features based on listeners' ratings (Vignoli & Pauws, 2005) or training a feature model on human annotated metadata such as genre (Aucouturier, Pachet, Roy, & Beurivé, 2007), or on perceptual similarity data (Novello et al., 2013).

Similarity is known to be context-dependent, which means that similarity can only be observed with respect to a certain context or characteristics (Berenzweig, Logan, Ellis, & Whitman, 2004; Cambouropoulos, 2009). Two pieces of music can be similar with respect to, for example, instrumentation, energy, or mode. Aspects of music similarity that have been studied separately include melodic similarity (Ahlbäck, 2007; Marsden, 2012; Müllensiefen & Frieler, 2004a; Orpen & Huron, 1992; Volk & van Kranenburg, 2012), rhythm similarity (Foote, Cooper, & Nam, 2002; Pohle, Schnitzer, Schedl, Knees, & Widmer, 2009), timbre similarity (Pachet & Aucouturier, 2004; Toiviainen et al., 1998), and harmonic similarity (de Haas, Rohrmeier, Veltkamp, & Wiering, 2009). These types of similarities have been termed dimensional similarity or sub-similarity.

The term context dependence has been used to denote the personal differences that may play a role in music similarity (Schedl & Knees, 2013). However, in this study, we focus on the fact that there is also agreement: certain levels of inter-rater agreement exist for music similarity (Jones, Downie, & Ehmann, 2007), and rhythm and timbre similarity (Honingh, Panteli, Brockmeier, López Mejía, & Sadakata, 2015).

While music similarity can generally be split into different sub-similarities, this does not mean that sub-similarities cannot overlap. For example, rhythm similarity can be influenced by timbre similarity. If there are two rhythmic patterns played by different instruments, it is the timbre of the instruments that makes the listener perceive two rhythmic patterns instead of one (Collins, 2006). Several models of rhythm similarity include timbre information (Lartillot, Eerola, Toiviainen, & Fornari, 2008; Panteli, Bogaards, & Honingh, 2014; Paulus & Klapuri, 2002).

To evaluate music similarity and sub-similarity models, several approaches have been taken. A distinction can be made between objective approaches, where genre, album, or artist information, for a database of music objects, is used to evaluate (sub-)similarity models (Aucouturier & Pachet, 2002; Logan & Salomon, 2001; Pampalk, Flexer, & Widmer, 2005), and subjective measures, where participants' ratings of perceived similarity serve as the golden standard (Novello et al., 2013). It should be clear that studies using objective evaluation make certain assumptions, for example, that genres have similar timbre or dance styles have similar rhythm. In addition to the aforementioned objective and subjective evaluations, simulations (i.e., composition of new rhythmic patterns) (Paulus & Klapuri, 2002), and non-musical sounds (i.e., synthesized tones) (Terasawa, Slaney, & Berger, 2005) have also been used. To our best knowledge, no sub-similarity model has yet been evaluated on perceptual data where music was rated.

In this study, we break the concept of similarity down into sub-similarities, and focus on rhythm and timbre similarity in Electronic Dance Music (EDM). The reason for restricting our study to EDM is that rhythm and timbre are prominent in this genre. They are the two most important musical dimensions (Reynolds, 2007, pp. 312–329). Moreover, we choose to focus on rhythm and timbre similarity of (mainly) polyphonic audio, whereas most studies use monophonic data. Although music similarity can be defined on pairs of entire songs, we can imagine that a piece of music is similar to only a part of another piece of music. This is particularly the case in genres like EDM, where a typical song exhibits several structural changes (Butler, 2006). Therefore, in this study, we focus on similarity of *segments* of music. We evaluate the models on perceptual data.

The remainder of this article is organized as follows. We will first give a brief introduction to EDM. In the section on "Segmenting music", the segmentation procedure that was used to segment the music is explained. Next, the section "Modeling similarity" presents the rhythm and timbre similarity models, and how they were trained and evaluated. We end with results, discussion, and a conclusion. Elements of the rhythm and timbre models presented in this article have been presented in previous papers (Panteli et al., 2014; Rocha, Bogaards, & Honingh, 2013).

Electronic Dance Music

EDM features electronically synthesized and sampled instrumentation, with at least some parts of a percussive nature, in tracks designed for dancing (Collins, Schedel, & Wilson, 2013, Chapter 8). Since its inception in the 1980s, many subgenres and hybrid genres have been introduced (Dayal & Ferrigno, 2013) at a rate unseen in any other genre (McLeod, 2001).[1] However, subgenre labels are not uniquely defined and may overlap (Collins et al., 2013, Chapter 8), often causing a musicological discourse. A general characterization can be given by dividing EDM into the "four on the floor" genres characterized by a 4-beat steady pattern, and the "breakbeat-driven" genres exhibiting more metrical irregularity (Butler, 2006).

Despite the breadth of styles, it is possible to name some consistent features of core EDM. An EDM track is often characterized by a steady tempo and a repeating bass-drum pattern (Butler, 2006). EDM has multiple complementary layers, some co-dependent in creating special effects, most capable of being independently dropped in and out (Collins et al., 2013, Chapter 8). Sampling, i.e., the practice in which a characteristic sound excerpt is recorded and mixed into a new piece of music, is also a frequent phenomenon.

Timbre and rhythm in EDM are often more important than melody and harmony (Reynolds, 2012). Timbre, especially, provides the cues by which music sections can be most easily distinguished (Yeston, 1976). In an EDM track, a change in timbre often consists of an instrument entering or leaving the mix. Rhythm, on the other hand, is typically based on a "loop", a repeating pattern of a particular, often percussive, instrument (Butler, 2006; Collins et al., 2013). A structural change in EDM is defined by the characteristic evolution of timbre and rhythm as opposed to a verse–chorus relation common in pop or folk music. Because an EDM track may exhibit several structural changes, similarity must be defined on the level of individual music segments.

Segmenting music

The segmentation of time series into meaningful, coherent units by automatically detecting their boundaries is a challenge crossing several scientific domains (Serrà, Müller, Grosche, & Arcos, 2012). A musical segment is a time-delimited section with some internal similarity or consistency in a given feature space, such as timbre or instrumentation. It therefore has temporal boundaries at its start and end (Casey et al., 2008). Tzanetakis and Cook (2000) stress the importance of segmentation in MIR: it is better to consider a song as a collection of distinct regions than as a whole with mixed statistics.

As argued earlier, timbre is a primary compositional parameter in EDM. Our segmentation method (Rocha et al., 2013) is based on the detection of timbral changes in the track over time, taking into account the idiosyncrasies of EDM. The first step of the algorithm is the detection of the first bass drum downbeat, as the entrance of the bass drum in an EDM track typically results in a decisive metrical representation (Butler, 2006). The first bass drum beat might not always align with the downbeat, but it usually introduces the beginning of a steady beat pattern, a prerequisite for tempo estimation and a rough estimate of a segment boundary. Tempo is then estimated in order to detect the duration of a beat, which is an important quantity because all subsequent features are extracted over frames, with frame lengths relative to the length of a beat. Next, a modified version of Foote's approach (2000) for novelty detection is applied on beat-aligned spectral magnitudes. At the end of this step, the algorithm yields a number of candidates for segment boundaries. The final step consists of dynamically aligning the obtained candidates with the closest downbeats, according to a set of musically informed heuristic rules (Rocha et al., 2013).

Modeling similarity

Like a number of other models that can be found in the literature (Aucouturier & Pachet, 2002; de Leon & Martinez, 2012; Dixon, Gouyon, & Widmer, 2004; Lartillot et al., 2008; Logan & Salomon, 2001), we model timbre and rhythm similarity by extracting low-level audio features from each music segment. We use the MIR Toolbox (Lartillot & Toiviainen, 2007) to extract most of the low-level descriptors considered in this study. For each segment, a feature vector is built that represents the timbre or rhythm aspects. To estimate similarity between two segments of music, the distance between the feature vectors corresponding to the segments is calculated.

Modeling 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 (Berber & Fales, 2005). These factors have been identified to include, amongst others, spectral flux, spectral centroid, and attack time (Burgoyne & McAdams, 2008; McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995; Peeters, Giordano, Susini, Misdariis, & McAdams, 2011). Most studies into this topic focused on monophonic timbres, but here we want to describe polyphonic textures. In this case we consider timbre as the perception of the polyphonic texture as a whole, also referred to as the "global sound" (Alluri & Toiviainen, 2010). Timbre in EDM can be characterized by a pattern of multiple instruments, for example, the combination of a kick sound, followed by a snare and a hi-hat sound that is repeated over time (as shown in Table 1). We assume that a timbre pattern is repeated consistently across the segment we analyze. To represent timbre in EDM, we focus on features that describe properties over the entire audible frequency spectrum. The following paragraphs describe the three types of features that we believe capture the dimensions of a polyphonic texture that are most relevant when comparing textures.

Mel-Frequency Cepstral Coefficients. Mel-Frequency Cepstral Coefficients (MFCCs) (Davis & Mermelstein, 1980) are used extensively in MIR algorithms to represent the spectral envelope of a given sound, which is one of the most salient components of timbre. The number of MFCCs that

Table 1. Inter-rater agreement measured in terms of ratings' correlations from timbre and rhythm similarity experiments.

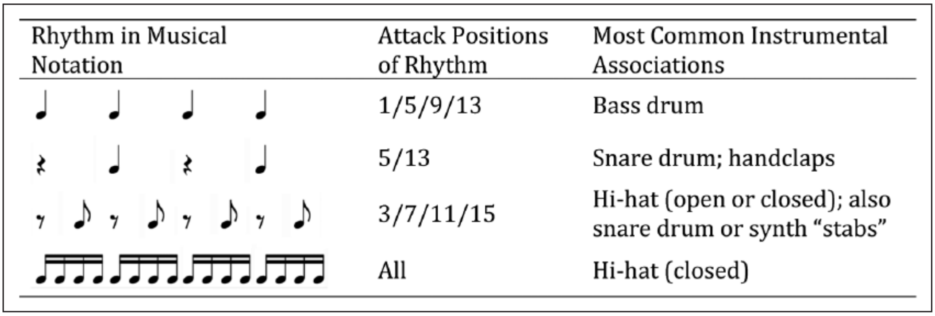
Raters' consistency	Rhythm	Timbre	Timbre*
Standard Deviation of correlations	0.16	0.12	0.14
Average correlation	0.64	0.51	0.52
VAF	**0.41**	**0.26**	**0.27**

Note. The VAF, calculated as the average correlation squared, is considered the upper bound for the performance of the models. Timbre* is a modified timbre dataset as described in the section "Regression results".

represent a spectral envelope well is open to discussion. Lower order MFCCs account for the broad shape of the spectral envelope, while the higher order ones describe the details of the spectrum (Aucouturier & Pachet, 2002). Therefore, while it is true that the more MFCCs we compute, the more precise the approximation of the signal's spectrum, a large number of MFCCs may not be beneficial, as we are only interested in the spectral envelope and not in the finer details of the spectrum (Aucouturier, Pachet, & Sandler, 2005). Aucouturier et al. (2005) report an optimum at 20 coefficients. We will use the same number in this work. We compute MFCCs in half-overlapping frames of 1-beat duration and take the average over all frames in time.

Spectral flatness. Facing the problem of audio matching (i.e., finding audio in a database that matches a given example), Herre, Allamanche, and Hellmuth (2001) searched for features that were both perceptually meaningful, and 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; 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 & Noll, 1984). 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 & Toiviainen, 2007). We compute spectral flatness in half-overlapping frames of 1-beat duration and take the average over all frames in time.

Dirtiness. The term "auditory roughness", also referred to as sensory dissonance, was introduced in the psychoacoustics literature by von Helmholtz and Ellis (1954). It is related to the phenomenon, called beating, which occurs whenever a pair of pure tones is close in frequency within a critical band, in a short period of time (Plomp & Levelt, 1965). It can be considered an attribute of timbre, as it relates to a signal's amplitude envelope and corresponding spectral distribution (Vassilakis & Kendall, 2010). The notion of roughness is approached from a different perspective in this work. Dirtiness is a term used by EDM listeners and producers when referring to a particular sound quality that is pervasive in 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 sound. Spectral analysis revealed that dirtiness might be (partly) explained by the detuning that producers apply to their synthesizer 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 sounding note, and can therefore be described using the concept of roughness (Vassilakis, 2001). We compute roughness in half-overlapping windows of 8 beats, as



Rhythm in Musical Notation	Attack Positions of Rhythm	Most Common Instrumental Associations
	1/5/9/13	Bass drum
	5/13	Snare drum; handclaps
	3/7/11/15	Hi-hat (open or closed); also snare drum or synth "stabs"
	All	Hi-hat (closed)

Figure 1. Example of a common EDM rhythm (Butler, 2006).

we believe dirtiness is represented by the roughness value over a large period of time, and take the average over all frames in time. We use the roughness estimation algorithm of Vassilakis (2001), which is a variant of the model by Sethares (1998).

The aforementioned features (20 MFCCs, 4 Spectral Flatness values, and 4 Dirtiness values) are combined in a feature vector that describes the timbre of a segment. The similarity between two different timbres is then given by calculating the distance (with an appropriate distance function) between the two associated feature vectors.

Modeling rhythm similarity

The study of rhythm and rhythm similarity has attracted attention from several multidisciplinary research domains. Research on rhythm perception includes studies to identify and quantify elementary rhythmic properties such as syncopation (Longuet-Higgins & Lee, 1984; Witek, Clarke, Wallentin, Kringelbach, & Vuust, 2014), complexity (Thul & Toussaint, 2008), and pulse clarity (Lartillot et al., 2008). In the MIR field, studies have focused on comparing rhythm sequences given in symbolic notation (Guastavino, Gómez, Toussaint, Marandola, & Gómez, 2009) or extracted from audio (Dixon et al., 2004; Paulus & Klapuri, 2002). Although rhythm information has been used in general music similarity modeling (Novello et al., 2013; Pampalk et al., 2005), no study to the best of our knowledge has explicitly addressed, and perceptually evaluated, rhythm similarity for EDM.

We model rhythm similarity of EDM excerpts using a wide range of musical and perceptual rhythmic attributes. Rhythm in EDM is constructed around the concept of a "loop", a repeating pattern associated with a particular (often percussive) instrument or instruments (Butler, 2006). As shown in Figure 1, a typical rhythmic pattern can consist of a combination of instrument sounds, thus exhibiting a certain rhythm polyphony. To analyze this, the signal is first split into rhythmic streams (analogous to "streams" as defined by Cambouropoulos, 2008). Then, for each stream, features are extracted based on three attributes: a) attack shape of the onsets, b) periodicity of rhythmic elements, and c) metrical distribution of events. For each segment, features are combined into a single feature vector by taking the average for each feature over separate streams, and pairwise similarity can be computed via suitable distance metrics. Details for each step are provided in the sections below.

Rhythmic streams. A rhythmic stream is defined as a single rhythmic pattern and may be produced by one or more instruments. Each stream contributes differently to rhythm perception (Griffiths & Warren, 2004). In order to describe rhythm in EDM, we must therefore distinguish between these streams. We apply an unsupervised method for detecting the number of streams and frequency range of each stream as follows: for a given audio segment, the FFT is computed using 46 ms windows with 3 ms hop size. The FFT magnitude spectra, transformed to a logarithmic scale to better approximate loudness (Klapuri, 1999), are then assigned to a total of 24 bark bands.[2] Synchronous masking is modeled using Schroeder's spreading function (Schroeder, Atal, & Hall, 1979) and temporal masking is modeled using a smoothing window of 50 ms (Painter & Spanias, 2000). The content of each bark band is then compared to every other band via the cosine distance, resulting in a similarity matrix between the 24 different bands. Correlating a checkerboard-like kernel along the main diagonal of the self-similarity matrix yields a novelty curve (Foote, 2003). The peaks of this novelty curve define the stream boundaries and indicate which (adjacent) bands should be grouped together to form a rhythmic stream.[3] Before summing the bark bands to the different rhythmic streams, however, an onset function is computed, as proposed by Klapuri, Eronen, and Astola (2006) and Scheirer (1998). The onset function we use is a weighted sum of magnitude spectra and their framewise difference. Onsets are then detected via peak extraction within each stream. The peak detection algorithm applied for both the novelty curve peaks and the onsets returns all local maxima above a threshold of 0.3 of the global maximum magnitude. This threshold was found optimal for onset detection in previous research (Panteli et al., 2014).

Attack shape. To distinguish between percussive and non-percussive patterns, features are extracted that characterize the attack shape of the onsets. Strictly speaking, this is foremost a timbre feature, but information about the percussiveness of the notes is expected to influence the perception of the underlying rhythm (Lartillot et al., 2008). Among other aspects, the attack time and attack slope are considered, as they are essential in modeling the perceived attack phase of onsets (Gordon, 1987). Attack time is the time in seconds taken for the onset function to rise from a local minimum to its respective local maximum (the selected onset). Attack slope is the gradient of the onset function between the local minimum and the onset. In general, onsets from percussive sounds have a short attack time and a steep attack slope, whereas non-percussive sounds have a longer attack time and a gradual slope. Similar to the attack slope is the attack shape sharpness, which represents the temporal derivative of the amplitude of onsets. Attack shape sharpness is different from the attack slope in that it represents the change in amplitude of (all) consecutive time frames of the onset function whereas attack slope focuses only on the average gradient of the attack phase of the selected onsets. The attack slope and attack shape sharpness were also used in modeling pulse clarity (Lartillot et al., 2008). For all onsets in all streams, the attack time, slope, and sharpness are extracted. Their means and standard deviations are combined into the feature vector.

Periodicity. One of the most characteristic style elements in the musical structure of EDM is repetition; the loop and consequently the rhythmic sequence(s), are repeating patterns. To analyze this repetition, the periodicity of the onset detection function per stream is computed using its autocorrelation and summed across all streams. The maximum delay taken into account is proportional to the bar duration. This is calculated assuming a steady tempo and $\frac{4}{4}$ meter throughout the EDM track (Butler, 2006). The tempo estimation algorithm of Rocha et al. (2013) is used. From the autocorrelation of the onset function we extract periodicity features (Panteli et al., 2014), such as lag duration of maximum autocorrelation, amplitude of maximum autocorrelation, harmonicity of peaks, and statistics of the autocorrelation distribution such as flatness, entropy, centroid, spread, skewness, and kurtosis.

Metrical distribution. To model the metrical aspects of the rhythmic pattern, the metrical profile (Smith, 2010) is extracted for each stream. First, the downbeat is detected as described by Panteli et al. (2014). Then, onsets per stream are quantized assuming a $\frac{4}{4}$ meter and 16th note resolution[4] (Butler, 2006) and the pattern is collapsed to a total of 4 bars. The latter is in agreement with the length of a musical phrase in EDM: usually a multiple of 4 (Butler, 2006). The metrical profile of a given stream is thus presented as a vector of 64 bins (4 bars × 4 beats × 4 sixteenth notes per beat) with real values ranging from 0 (no onset) to 1 (maximum onset amplitude). As in the above section "Periodicity", we assume a steady tempo, $\frac{4}{4}$ meter, and a unique repeated rhythmic pattern throughout the EDM segment, for both downbeat detection and onset quantization. For each rhythmic stream, a metrical profile is computed and, from it, the following features are extracted: syncopation, symmetry, density, fullness, and center of gravity (Panteli et al., 2014). Features are computed per stream and averaged across all streams.

Aside from these features, the metrical profile itself is also added to the final feature vector as a multidimensional feature. This was found to improve results in related research (Smith, 2010). In the current approach, the metrical profile is provided for the bottom 4 streams only. We decided to restrict the metrical profile to 4 streams after empirical observation with 4 subjects on a set of 120 characteristic songs in our dataset (Panteli et al., 2014). We found that rhythm could be described by an average of 4 ($M = 3.63$, $SD = 0.27$) rhythmic streams. Moreover, the stream detection algorithm returned on average $M = 4.49$ streams with standard deviation $SD = 0.99$. If more than 4 streams are detected, the remaining top streams are ignored in the computation of the metrical profile as they correspond to high frequency content often not as essential to defining the rhythmic pattern as the bottom 4 streams. If fewer than 4 streams are detected we pad the remaining metrical profile values with zeros. This results in a total of 64 × 4 = 256 values.

Evaluation

To evaluate the models of timbre and rhythm similarity, perceptual data from a previous study (Honingh et al., 2015) were used. Audio features and perceptual data were combined in a regression model for rhythm and timbre similarity. Linear regression is a common strategy to describe the relationship between predictor and response variables. It has been used in several studies of music similarity (Cui, Shen, Cong, Shen, & Yu, 2006; Müllensiefen & Frieler, 2004b; Novello et al., 2013). The regression returns a weight for each of the features. Additional and more complex models for learning feature weights (McFee, Barrington, & Lanckriet, 2010; Stober & Nürnberger, 2013; Wolff & Weyde, 2014) can be investigated in future research.

Before applying regression, features were preprocessed, and feature selection was performed. We checked for non-linearity in the data by applying non-linear transformations (square, cube, square root, log) to the input features as mentioned in Novello et al. (2013). These non-linear transformations did not improve the accuracy of the model, which supports the assumption that a linear model is appropriate for our data.

Perceptual data from experiment

In a previous study (Honingh et al., 2015), perceptual data on timbre and rhythm similarity were collected. Participants were invited to listen to pairs of musical segments and rate the timbre/rhythm similarity on a 4-point scale with levels: 1) dissimilar, 2) somewhat dissimilar, 3) somewhat similar, and 4) similar. Two of these experiments were performed: one for rhythm similarity and one for timbre similarity.

Following previous experiments on music similarity (Jones et al., 2007; Novello, McKinney, & Kohlrausch, 2011), we did not explain the concept of similarity to avoid possible bias. Instead, participants were presented with synthesized example pairs at the beginning of each experiment representing the concept of timbre and rhythm similarity, respectively. In the rating phase, participants were asked to listen to pairs of EDM segments ("Listen to the following audio clips") and to rate the pairs based on the timbre/rhythm similarity of the segments ("How similar are they in their timbre/rhythm?").

Each participant rated a total of 60 pairs of segments. The order in which pairs were presented was randomized every time. At the end of the experiment participants were asked to complete a questionnaire to give us more information on their age, gender, and expertise in music in general and in EDM in particular.

The stimuli for this experiment were a balanced selection of twenty 12-second segments from a database of commercially released EDM tracks (see Appendix, Table A1). A pilot experiment was conducted following the above procedure to cross-check the choice of the music material. To make sure that rhythm similarity could be measured relatively independently from timbre similarity (and the other way around), pairs of music segments were made in the following four categories, based on the pilot experiment:

- high rhythm similarity – high timbre similarity
- high rhythm similarity – low timbre similarity
- low rhythm similarity – high timbre similarity
- low rhythm similarity – low timbre similarity

The stimuli for the experiment were further constrained in their genre and tempo, since both these parameters are important when rating music similarity (Novello et al., 2011). We controlled for genre by restricting the dataset to EDM, and consulting an EDM expert to help us balance the presence of various subgenres in the dataset. Tempo is an important factor in music similarity too: people tend to rate pieces that have a similar tempo as similar overall. The music stimuli were therefore also restricted to the 110–133 bpm tempo range. This tempo range is centered around 120 bpm, a default tempo setting for several digital audio workstations used in music production (Nash & Blackwell, 2011), and a preferred tempo in the perception of human listeners (van Noorden & Moelants, 1999).

The experiment on timbre similarity was completed by 62 participants, 31 female and 31 male, between 19 and 65 years old ($M = 27.1$, $SD = 9.1$). More than half of the participants (40) had received formal musical training between the ages of 5 and 25 in music styles including classical, pop, rock, and jazz. A total of 21 participants stated that they work with music professionally, mainly as producers and audio engineers. Most of the participants (55) stated familiarity with EDM to variable degrees but they reported knowing on average less than 20% of the stimuli of the experiment. The rhythm similarity experiment was completed by 57 participants, 13 female and 44 male, between 17 and 52 years old ($M = 27.9$, $SD = 8.0$). A total of 31 participants had received musical training between the ages of 4 and 20 in music styles including classical, jazz, pop, rock, and electronic. Only 19 participants reported that they work with music professionally, mainly as musicians. Although 51 participants were familiar with EDM to variable degrees, they reported knowing less than 20% of the stimuli on average.

The ratings from all participants resulted in two similarity matrices, one for rhythm and one for timbre, in which every pair of music segments was rated by at least 10 participants. A cross-participant concordance check was performed. The average ratings for each pair are used for training and testing the models in the present study.

Method

As explained earlier, the perceptual ratings are used to train and evaluate a linear model of similarity predictions. For each pair of segments, the audio features were combined to provide a similarity estimate of the corresponding perceptual rating. The following sections give detailed information of this process, including the necessary feature preprocessing, feature selection, cross-validation, and linear regression.

Feature preprocessing

Linear regression assumes normality of the distribution of the input data. To ensure this we considered preprocessing the feature data.

Our feature space consisted of 21 rhythm features and 28 timbre features computed for the 20 audio segments used in the corresponding perceptual experiments. For each feature we scale the values to the [0, 1] range. We then test for normality using the Shapiro Wilk test (Shapiro & Wilk, 1965). Features with non-normal distribution are transformed using a logarithmic transformation:

$$\bar{X} = \log\left(X + 1\right) \tag{1}$$

where X is the normalized sample distribution of a given feature and $\bar{X}$ the distribution after the transformation. The addition of 1 is used to avoid infinite and negative log values. Applying the Shapiro Wilk again test after the logarithmic transformation showed normality for all features.

The computational features described above in the section "Modeling similarity" are extracted for every audio segment. The perceptual similarity ratings are collected for every pair of segments. To be able to combine computational features and perceptual ratings we had to extract feature representations for every pair of segments. Following the approach by Novello et al. (2013), we model the computational distance of a pair of segments by taking the absolute difference of the feature values corresponding to the underlying segments of the pair. Here we interpret the rhythm/timbre similarity of a pair of segments as the inverse distance between these segments; i.e., high similarity corresponds to small distance between two segments.

For example, let s_a and s_b define the two audio segments of pair *i*, where *i* runs over all combinations of segments *a* and *b* with *a*≠*b*. If $X_{s_a} = \left(x_{s_a,1},\ldots,x_{s_a,N}\right)$ and $X_{s_b} = \left(x_{s_b,1},\ldots,x_{s_b,N}\right)$ define the N features for segments *a* and *b*, respectively, the distance between the segments of pair *i* is given by

$$X_i = \left|X_{s_a} - X_{s_b}\right| = \left(\left|x_{s_a,1} - x_{s_b,1}\right|,\ldots,\left|x_{s_a,N} - x_{s_b,N}\right|\right). \tag{2}$$

All features were processed according to Equation 2 above, except for the metrical profile rhythm feature (see section "Metrical distribution"). The metrical profile is a multidimensional feature consisting of 64 × 4 = 256 values. To estimate the distance between the metrical profiles of two segments, we used the Euclidean distance.[5] Let $\boldsymbol{x}_{s_a,m}$ and $\boldsymbol{x}_{s_b,m}$ represent the metrical profiles for segments s_a and s_b, respectively, where $m = m_1,\ldots,m_J$ for J = 256 values. The Euclidean distance of two metrical profiles is defined as

$$d\left(\boldsymbol{x}_{s_a,m}, \boldsymbol{x}_{s_b,m}\right) = \sum_{j=1}^{J} \sqrt{\left|x_{s_a,m_j} - x_{s_b,m_j}\right|^2}. \tag{3}$$

Therefore, for the rhythm model only, the feature values for pair *i*, expressing the distance between the segments *a* and *b*, were obtained by the feature differences of the one-dimensional features as defined in Equation 2 and the Euclidean distance of the metrical profiles as defined in Equation 3. This is denoted as

$$X_i = \left(\left|x_{s_a,1} - x_{s_b,1}\right|, \ldots, \left|x_{s_a,N} - x_{s_{b,N}}\right|, d\left(\boldsymbol{x}_{s_a,m}, \boldsymbol{x}_{s_b,m}\right)\right). \tag{4}$$

The timbre model has no such multidimensional feature as the metrical profile, therefore the distance between the segments of pair *i* is given by Equation 2. To be able to combine feature differences and similarity ratings, the similarity ratings were inverted (i.e., transforming "similarity" to "distance") and scaled to fit into the range of [0, 1]. A rating of value 0 denotes small distance and hence high similarity and a rating of value 1 denotes large distance and hence low similarity.

Feature selection

In the listening experiments, a total of 190 pairs of segments were used. Therefore, for each of the models (timbre and rhythm), 190 data points could be used for training and testing. As the total number of rhythm and timbre features (21 and 28, respectively) was relatively large compared to the number of data points, we performed feature selection in order to reduce the number of features. We chose a correlation-based feature selection algorithm but other dimensionality reduction techniques such as Principal Component Analysis or Non-Negative Matrix Factorization (Sun, Ji, & Ye, 2013) could be applied.

The correlation-based feature selector proposed by Hall (1999) and implemented in the Weka[6] framework was used. This algorithm ranks feature subsets according to a correlation based heuristic evaluation function. The evaluation function selects subsets of features highly correlated with the ground truth label (also referred to as "class") and uncorrelated with each other. This is based on the heuristic "merit", M_F, of a feature subset *F* estimated with the following equation:

$$M_F = \frac{k r_{fc}}{\sqrt{k + k(k-1) r_{ff}}} \tag{5}$$

where *k* is the number of features contained in subset *F*, r_{fc} is the mean feature-class correlation and r_{ff} is the mean feature-feature intercorrelation. The numerator of this equation indicates how well the set of features predicts the class whereas the denominator indicates how much redundancy there is among the features (Hall, 1999). We used the above feature selector with the best-first search algorithm and a stopping criterion that terminates the search after five consecutive fully expanded subsets show no improvements over the current best subset.

Cross-validation

To optimize prediction and limit overfitting effects of linear regression we used cross-validation. The perceptual ratings were split into two non-overlapping sets, training and test set. The training set contained the known data on which the model was trained and the parameters were learned, and the test set consisted of the unknown data on which the model was tested. The performance of the model on the test set indicates how well the model could generalize to unknown datasets.

Our dataset consisted of 190 pairs resulting from pairwise combinations of a total of 20 segments. For optimal results we aimed to train and test the model with sets of non-overlapping pairs, and consequently, non-overlapping segments. We used leave-one-out cross-validation where one pair was kept for testing and the remaining (non-overlapping) pairs were used for training the model. That is, 1 pair (i.e., 2 out of 20 segments) was held for testing, and 153 pairs (i.e., $\binom{18}{2} = 153$ pairs resulting from the remaining 18 segments) were used for training. We applied this training-test split for each of the 190 pairs. The outcome of cross-validation was averaged to provide the end result.

Linear regression

We used linear regression to estimate the ability of our features to approximate the perceptual ratings of timbre and rhythm similarity. Using cross-validation as described above, the model was trained with features and similarity ratings from 153 pairs and tested by predicting the similarity of a single (non-overlapping) pair. This process was repeated for each pair. Linear regression uses the least squares method to learn coefficients that best fit the data. In our case, the least squares method minimizes the difference between similarity ratings provided by the listeners and similarity predictions estimated via the model.

To formally describe the above process, let $y = y_1,\ldots,y_I$ denote the similarity ratings provided by the participants, and $\hat{\boldsymbol{y}} = \widehat{y_1},\ldots,\widehat{y_I}$ denote the similarity estimates provided by the model for a total number of *I* pairs. Let $\beta = \beta_0, \beta_1, \ldots, \beta_N$ denote the coefficients that are linearly combined with feature variables $X_i = x_{i_1},\ldots,x_{i_N}$ for pair *i* to provide the similarity estimate of the model, then

$$\widehat{y_i} = \beta_0 + \beta_1 x_{i_1} + \ldots + \beta_N x_{i_N}. \tag{6}$$

With linear regression we find coefficients β that minimize the sum *S* of squared residuals given by

$$S = \sum_{i=1}^{I} (y_i - \widehat{y_i})^2. \tag{7}$$

In Equations 6 and 7 of linear regression, y_i denotes the dependent variable and X_i the independent variables.

Using cross-validation we applied linear regression for each of our training-test sets. The β coefficient values from each regression experiment were averaged to provide the overall result denoted "average β". These coefficients indicate the importance of the selected features for modeling timbre and rhythm similarity. To evaluate the overall performance of the cross-validated linear regression we computed the Pearson's correlation *r* between predicted distances $\hat{\boldsymbol{y}}$ of the testing sets and perceptual similarity ratings *y*. We report the accuracy of the model in units of the Variance Accounted For (VAF) defined as

$$VAF = correlation\left(y, \hat{y}\right)^2. \tag{8}$$

In results we also report the absolute error of predicted $\hat{\boldsymbol{y}}$ versus true y ratings averaged across all training-test sets, and the root mean squared error.

Results

In this section we present results for the rhythm and timbre similarity models. First, we present the upper bound of the performance of each model estimated via the inter-rater agreement of the perceptual ratings as described in the section "Upper bound of model performance". Results from the regression analysis are compared to the upper bound to check whether optimal performance has been achieved. The linear regression coefficients, interpreted as individual weights of the model's parameters, indicate the importance of the selected audio features and the degree to which each of them contributes to perceptual timbre or rhythm similarity.

Upper bound of model performance

For subjective tasks like assessing music similarity, ratings can vary from person to person. When modeling music similarity based on human ratings there exists an upper bound of how well the model could (or should) perform. This depends on the inter-rater agreement of the participants (Flexer, 2014; Novello et al., 2013).

There are several ways to assess the inter-rater agreement. A common approach is to calculate the mean of pairwise correlations among raters. In our listening experiment, every participant rated a random subset of 60 pairs from a total of 190 pairs (Honingh et al., 2015). Since participants rated different sets of pairs we couldn't measure pairwise correlations among all raters. Instead, for each pair we calculated the average of its ratings. Then for each participant we computed the Pearson's correlation between his/her ratings and the average ratings of the corresponding pairs. We repeated this for all participants and computed the average of all Pearson's correlations. Following Equation 8, we squared this average correlation to compute the VAF. This defined the overall agreement of the ratings and the optimal performance we could expect from our model. The upper bound results for the rhythm and timbre similarity models are presented in the paragraph below. More details on how we assess the inter-rater agreement can be found in the related study (Honingh et al., 2015).

For rhythm similarity we collected ratings from a total of 57 participants. Following the procedure described above, the mean, standard deviation, and VAF of participants' correlations are summarized in Table 1. The upper bound for the rhythm similarity model was VAF = 0.41. For timbre similarity we considered ratings from 23 participants.[7] The agreement on timbre similarity ratings was not very high (slight agreement on the Landis and Koch scale), but it was, however, comparable to the agreement on rhythm similarity and general music similarity (Honingh et al., 2015). The upper bound for the timbre similarity model was VAF = 0.26. We further report the upper bound for timbre similarity after eliminating two outlier segments from the dataset as discussed in the next section, "Regression results". The upper bound for this modified dataset was VAF = 0.27.

Regression results

For each model (timbre and rhythm) we ran linear regression with the cross-validation method as described in the above "Cross-validation" section. Regression accuracy was measured in terms of VAF as described in the "Linear regression" section. The results for timbre and rhythm similarity models are summarized in Table 2 and explained in the sections below. The β coefficients estimated in each regression analysis correspond to weights for the individual audio features. As described earlier, these coefficients were averaged across the 190 regression runs

Table 2. Results from cross-validated linear regression for timbre and rhythm similarity models.

	Rhythm	Timbre	Timbre*
Mean absolute error	0.16	0.17	0.16
Root mean squared error	0.20	0.21	0.19
Correlation coefficient	$r = .62, p < .05$	$r =. 18, p < .05$	$r = .32, p < .05$
VAF	0.38	0.04	0.10

Note. Timbre* corresponds to a modified timbre dataset as described in section "Regression results".

Table 3. Average β weights for rhythm features as resulted from cross-validated linear regression.

Average β	Feature	Category
0.61	autocorrelation skewness	Periodicity
0.38	autocorrelation maximum amplitude	Periodicity
0.31	autocorrelation centroid	Periodicity
0.17	attack shape sharpness SD	Attack Shape
0.15	weighted syncopation	Metrical Distribution
0.07	harmonicity of autocorrelation peaks	Periodicity
0.06	symmetry of metrical profile	Metrical Distribution

Table 4. Average β weights for timbre features as resulted from cross-validated linear regression.

Average β	Feature	Category
0.27	MFCC 8	MFCCs
0.26	flatness band 1	Spectral Flatness
0.21	dirtiness band 1	Dirtiness
0.13	MFCC 6	MFCCs
0.07	MFCC 3	MFCCs
0.07	MFCC 2	MFCCs
0.03	MFCC 4	MFCCs

to provide the final weights of the features. We summarize the average β coefficients in Table 3 and Table 4 for rhythm and timbre similarity models, respectively.

The rhythm model achieved an accuracy of 0.38. This reached close to the upper bound (0.41) for rhythm similarity, which indicates that optimal performance for the given input data was achieved. For the timbre model, the accuracy (initially) achieved was 0.04 for the upper bound of 0.26. We investigated the reason for this relatively low result by looking at the input dataset for any possible outliers. We found that 2 out of the 20 segments used in the experiment had relatively distinct timbre with respect to the rest of the segments. We removed these from the timbre dataset and reapplied regression analysis. This increased the accuracy to 0.10 for the upper bound of 0.27. We refer to this modified timbre dataset as *Timbre**. In Table 2 we report results of timbre similarity for both before eliminating these segments from the dataset, i.e., dataset Timbre, and after, i.e., dataset Timbre*.

Results for rhythm model. For the rhythm model a total of 7 features were found to be important for modeling similarity (see Table 3). The majority (4 out of 7) of these features belong to the

Periodicity category. This may emphasize the importance of repetition in modeling rhythm similarity, especially for music styles like EDM. From the remaining features, 2 come from the Metrical Distribution category and 1 from the Attack Shape category. This distribution of features suggests that all three categories contribute to rhythm similarity.

Focusing on the individual coefficients of the rhythm features we can observe differences of the relative importance of each feature. First, the skewness of the autocorrelation curve was weighted with average β of 0.61 which was notably more than the weight of the second important feature (average β of 0.38). Having a skewed distribution of the autocorrelation curve means having strong periodicities of a certain level, i.e., rhythmic events strongly repeating at the bar level but not at the beat level and vice versa. Distinguishing between such distributions seems to be important for rhythm similarity in EDM. The second group of features consisted of the maximum autocorrelation value and the centroid of the distribution with average β coefficients of 0.38 and 0.31, respectively. The former feature describes how strong the periodicity of the onset function is and the latter describes the geometric center of the distribution which, similar to skewness, indicates the important levels of periodicity. The fourth feature in the ranking was the standard deviation of the attack shape sharpness followed by the weighted syncopation value with average β values of 0.17 and 0.15, respectively. Sharpness is another way of measuring percussiveness of an instrument's onset and its standard deviation measures the variation of percussive or non-percussive onsets. The weighted syncopation measures syncopation within each stream, i.e., within each rhythmic pattern, and adds the syncopation values by weighing each stream differently. The last group of features consisted of the harmonicity of autocorrelation peaks with average β of 0.07 and the symmetry of the rhythmic pattern with average β of 0.06. The harmonicity feature measures whether periodicities occur at harmonic relations with respect to the beat, i.e., periodicity at beat sub-divisions and multiples. Symmetry describes whether the rhythmic pattern is the same for each of the 4 bars as summarized in the metrical profile. Overall, the degree and number of layers of repetitions, together with syncopation and percussiveness of the instrument were found important for rhythm similarity in EDM.

To further investigate the importance of each feature category we performed a separate regression analysis for each feature subset, Attack Shape, Periodicity, and Metrical Distribution. Figure 2a shows results for a) a selection of features from the correlation-based feature selector (the subset that obtained the optimal performance reported in Table 2), b) all features, c) features from the Attack Shape, iv) from the Periodicity, and v) Metrical Distribution of rhythmic events. It was observed that Periodicity features outperformed the predicting ability of features from the other two categories (a finding also supported in Panteli et al., 2014). Moreover, the performance of features from only the Periodicity category reached close to the performance of features selected by the correlation-based method described in the section "Feature selection". Features from the Attack Shape and Metrical Distribution categories performed poorly compared to features from the Periodicity category.

Results for timbre model. For the timbre model, the interpretation of the features is less straightforward, since the model achieved low accuracy on the full Timbre dataset and moderate accuracy on the adjusted Timbre* dataset. In the paragraphs below we report our findings, but we have to keep in mind that these may not generalize entirely to other EDM datasets.

The timbre features selected by the correlation-based algorithm (in section "Feature selection") are listed in Table 4. These consisted of 1 Spectral Flatness feature, 1 Dirtiness feature and 5 MFCCs. The MFCCs dominated the selected feature subset (5 MFCCs out of a total of 7 features).

Focusing on the ranking of the features we observe that the 8th MFCC, flatness of band 1, and dirtiness of band 1 occupied the first places with average β coefficients of 0.27, 0.26, and

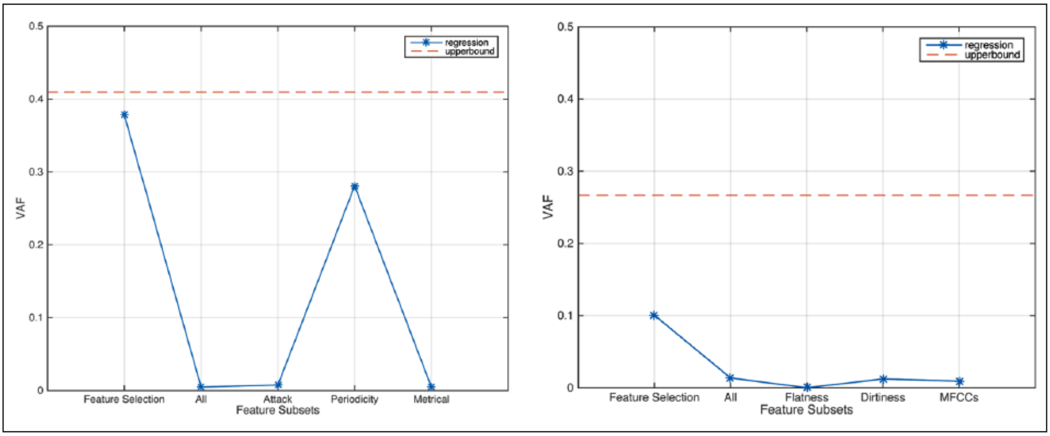


Figure 2. VAF from regression analysis of different subsets of features. For the timbre model (right figure) results correspond to the dataset after eliminating the two outlier segments.

0.21, respectively. The MFCCs preserve details of the spectral shape and in this case the amount of details captured by the 8th MFCC in particular was shown to be important for timbre similarity. The flatness in the first frequency band (0–200 Hz) describes the energy spread in the low frequencies. Likewise, the dirtiness in the first frequency band indicates the roughness in the low frequencies. This indicates that for timbre similarity in EDM it is important whether the low frequency instruments (e.g., kick or bass drum) are present in the mix or not, and the way they sound, i.e., being rough or not. The next features in the ranking corresponded to 4 features from the MFCCs category, namely the 6th, 3rd, 2nd, and 4th MFCCs. The 6th MFCC with average β of 0.13 preserves also some details of the spectrum, though relatively less fine-grained than the 8th MFCC. The remaining features, 3rd, 2nd, and 4th MFCCs respectively, obtained average β coefficients lower than 0.07. These features represent the spectral shape at a more global than detailed level. Overall, the presence and quality of the low frequency sounds as well as descriptors of the global spectral shape of the instruments' mix were shown to be important for timbre similarity in EDM.

As in the rhythm case we investigated the importance of each feature category via performing regression analysis for each feature subset, Spectral Flatness, Dirtiness, MFCCs. Figure 2b shows results for a) a selection of features from the correlation-based feature selector (the subset that obtained the optimal performance reported in Table 2), b) all features, c) Spectral Flatness, iv) Dirtiness, and v) MFCCs. It is observed that features of the Dirtiness category outperformed the features of the MFCCs and Spectral Flatness categories. Furthermore, the feature subset selected via the algorithm described in the section “Feature selection”, obtained a much higher performance than the individual feature categories.

Discussion

We developed computational models for timbre and rhythm similarity and evaluated them with perceptual ratings. The evaluation of the feature model for rhythm similarity has shown that good predictions, close to the upper bound of the task, could be made. For the model on timbre similarity, only moderate accuracy could be achieved. In the paragraphs below we review the process of modeling timbre and rhythm similarity, provide possible explanations for the low accuracy of the timbre similarity model, and suggest directions of improvement for future work.

We considered an initial set of 28 timbre features and 21 rhythm features to model timbre and rhythm similarity, respectively. These features were chosen based on their state-of-the-art performance and relevance. Due to the limited number of song pairs analyzed via the experiment (190 pairs) and the relatively large number of features (28 for timbre and 21 for rhythm) it was necessary to reduce the dimension of the feature space for more robust results.

We used a correlation-based feature selection algorithm (in section "Feature selection") to automatically select features most correlated with the annotations and least correlated with each other. We ran the feature selector on several feature subsets with randomized feature order. Amongst the best performing feature subsets, there were between 7 and 10 features selected in total, and most of these features overlapped. We chose the feature subset with the highest feature correlation merit. Different feature selection or dimensionality reduction techniques could be considered in future work.

The features shown important for rhythm similarity in EDM could be grouped in three broad musical categories, a) repetition (autocorrelation skewness, maximum amplitude, centroid), b) percussiveness (attack shape sharpness SD), and c) metrical hierarchy (weighted syncopation, harmonicity of autocorrelation peaks, symmetry of metrical profile). In the first category, autocorrelation maximum amplitude indicates the degree of repetition whereas skewness and centroid indicate the metrical levels at which this repetition occurs. For high similarity between two segments the rate at which patterns repeat as well as the location they appear within a bar should be consistent. From the second category, percussiveness, we note that similarity is better achieved when the timbre of the instruments performing the rhythmic patterns is consistent. The third category, metrical hierarchy, hints that syncopation or metrical structure should be consistent for two segments to be perceived as similar in their rhythm.

The results on the importance of features for rhythm similarity (Table 3) are in agreement with other studies on rhythm. First, the importance of periodicity for describing rhythm has also been supported in various studies where features describing periodicity of rhythmic elements are successfully applied for rhythm description, classification, or general similarity (Dixon et al., 2004; Lartillot et al., 2008; Novello et al., 2013; Paulus & Klapuri, 2002). Moreover, features related to periodicity of the onset detection function are highly correlated with perceptual ratings of pulse clarity (Lartillot et al., 2008). Fluctuation patterns (Pampalk et al., 2005), a periodicity-based feature, are also used in the best performing algorithms of the MIREX Audio Music Similarity task (Pohle & Schnitzer, 2009; Seyerlehner, Schedl, Knees, & Sonnleitner, 2011), as well as the general music similarity in the analysis of Novello et al. (2013). In the latter, attack shape features are also ranked amongst the best rhythmic features for the task of general music similarity. Syncopation is also important for modeling rhythm (Longuet-Higgins & Lee, 1984; Smith, 2010). Moreover, the concept of weighted syncopation shares similarity with the syncopation measure defined by Witek et al. (2014), where instrumental weights are added to account for the polyphonic nature of drum breaks.

The features selected for timbre similarity consist of 5 MFCCs, and flatness and dirtiness of the low frequencies (0–200 Hz). MFCCs capture details of the spectral envelope whereas dirtiness and flatness capture aspects of timbre roughness and spectral spread, respectively. The fact that MFCCs dominated the selected feature subset supports the importance of MFCCs in timbre content description as indicated in the literature (Aucouturier et al., 2005). However, MFCCs achieved low performance when correlated with perceptual ratings of timbre similarity. One feature that was not mentioned in previous research and was shown to be important in our model is "dirtiness". Dirtiness is a term used by EDM listeners and producers when referring to a particular sound quality that is pervasive in synthesizers. Analysis showed that dirtiness of especially the low frequency instruments is important for timbre similarity in EDM.

It is difficult to say whether there is consensus on the features used for timbre similarity compared to other studies on timbre. The first reason for this is because we have considered polyphonic timbre compared to the monophonic timbres used in previous studies (Grey, 1977; McAdams et al., 1995). Second, we have used a dataset with a specific range of sounds (EDM), which does not generalize to all sounds and timbres. Third, we chose to evaluate our timbre model with perceptual ratings of similarity whereas other studies use ground truth sets based on genre labels (Pachet & Aucouturier, 2004). In this light our ground truth set exhibits ratings that are more fine-grained than the more general timbre boundaries of genres.

The relatively low correlation between timbre features and perceptual ratings could be explained by multiple factors. First, the set of timbre features could be improved and expanded. We based our analysis on an initial set of 28 timbre features (in section "Modeling timbre similarity") that was later reduced to 7 timbre features (in section "Feature selection"). However, different features sets might be considered. For example, the Timbre Toolbox (Peeters et al., 2011) could be considered as an initial timbre feature set and a different feature selection might yield better results. While the chosen features capture some timbral aspects of a polyphonic musical excerpt, different descriptors could be considered to represent relations and interactions of the multiple musical instruments in the mix. Timbre descriptors are calculated on a short (sometimes less than a second) frame base while timbre perception might depend on longer temporal windows. In future research, features preserving more time information can be considered.

Another reason that only moderately successful predictions could be made with the timbre similarity model is possibly that the dataset was too small for the complex and multidimensional concept of timbre similarity. Rating timbre similarity turned out to be a difficult task and some participants were shown to use rhythm features to rate timbre similarity (Honingh et al., 2015). Although these participants were removed from the dataset[8] it is possible that the dataset still contained ratings from participants who used other than timbre features to rate timbre similarity. This could also be an indication that the concepts of timbre and rhythm cannot be completely separated. If indeed the perceptual ratings were not solely based on timbre features, the model that was exclusively based on timbre features could never have made accurate predictions of these data. The low accuracy of timbre similarity compared to rhythm similarity could also be due to timbre being more multifaceted than rhythm. We cannot conclude this from our current data but more evidence could be gathered in future work.

Various strategies for the estimation of the upper bound of model performance have been mentioned (Flexer, 2014; Novello et al., 2013). In our case, the upper bound was estimated as the VAF from the average correlation between participants and the "average participant", i.e., the average of the ratings of a corresponding pair. This approach compensated for missing data. The upper bound reported by Novello et al. (2013) for the task of general music similarity was 0.45. This was relatively close to the upper bound of 0.41 for the rhythm similarity task in our study. The upper bound for the timbre similarity task was lower, 0.24, possibly as a result of different rating strategies, population sub-groups, and difficulty of the task of rating timbre similarity (Honingh et al., 2015).

Different listener strategies may exist when rating timbre or rhythm similarity between pairs of segments. It was shown that musically trained people rated rhythm similarity significantly differently than musically untrained people, and that people who are familiar with EDM rated timbre similarity significantly differently than people who are not familiar with EDM (Honingh et al., 2015). In this study, we chose to use the full dataset and as such to optimize the model for an average listener. However, the model could be optimized for different groups of people in future research.

Conclusion

This study focused on modeling rhythm and timbre similarity in EDM. We assumed rhythm to be different from timbre and modeled rhythm and timbre similarity separately. We focused on the concept of sub-similarity, in contrast to other studies focusing on general music similarity, which allowed us to look with more details into the underlying dimensions and the parameters they define. Our evaluation method relied on similarity captured via perceptual ratings rather than arbitrary genre labels. From state-of-the-art research in music information retrieval and musicology, we selected and designed audio descriptors that capture aspects of timbre and rhythm in EDM. These features were evaluated with perceptual ratings of similarity collected via a listening experiment.

We applied a linear regression model to correlate audio descriptors with perceptual data and assess the importance of each descriptor in rating timbre and rhythm similarity. The accuracy of the rhythm model reached close to the upper bound of model performance. The accuracy of the timbre model was relatively low possibly due to data not being very representative of the multifaceted concept of timbre. Features related to the periodicity of rhythmic elements were amongst the most important descriptors for rhythm similarity in EDM. In particular, the degree of repetition as well as the metrical position of the repeating pattern were found highly correlated with perceptual ratings. Features related to the presence or absence of low frequency instruments in the mix and the roughness in their sound were found to be important for modeling timbre similarity in EDM.

We believe that future research focusing on sub-similarity in a similar manner could contribute significantly to understanding the complex concept of music similarity. Both audio descriptors as well as perceptual ratings of similarity as presented in this study could be improved and expanded. With sufficient data one could then work towards modeling the interaction between rhythm and timbre similarity. The concept of general music similarity could be further investigated in terms of the weighted combination of several sub-similarities including melody, harmony, timbre, and rhythm. Findings from related research could contribute to the development of a music similarity model beneficial for musicology, music perception, and music information retrieval.

Funding

Aline Honingh has been supported by an NWO-VENI grant 639.021.126. Maria Panteli and Bruno Rocha have been supported by a grant from the Centre for Digital Humanities, Amsterdam.

Notes

1. The terms “subgenre” and “style” will be used interchangeably here, in order to cite the used resources as closely as possible. With both terms we mean the stylistic attributes of the music without the cultural context.
2. Bark bands are preferred to other filter banks, as they have been used in several psychoacoustic models of loudness and masking including the ones considered in this study (i.e., Schroeder’s spreading function).
3. In this case it is desirable to group contiguous bands as a way to represent broad frequency ranges (i.e., “the lows, mids, and highs”).
4. 16th notes in swing style are quantized to sensible positions in this algorithm.
5. The hamming and cosine distance of the metrical profiles were also tested but showed lower performance. Other measures considering metrical hierarchy (Guastavino et al., 2009) can be tested in future work.
6. See http://www.cs.waikato.ac.nz/ml/weka/index.html
7. We considered only a subset of the total number of participants who rated timbre similarity for reasons considered in the “Discussion” section and in Honingh et al. (2015).

8. It was found that a number of people who were asked to rate timbre similarity, used rhythm characteristics to do so (Honingh et al., 2015). The data from this group of people were left out of the dataset that we used in this study for the evaluation of the timbre similarity model.

References

Ahlbäck, S. (2007). Melodic similarity as a determinant of melody structure. *Musicae Scientiae, 11*(1), 235–280.

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

Aucouturier, J.-J., & Pachet, F. (2002). Music similarity measures: What's the use? In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)*. Retrieved from http://ismir2002.ismir.net/proceedings/02-FP05-2.pdf

Aucouturier, J.-J., Pachet, F., Roy, P., & Beurivé, A. (2007). Signal + context = better classification. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 425–430). Retrieved from http://ismir2007.ismir.net/proceedings/ISMIR2007_p425_aucouturier.pdf

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.

Berber, H. M., & Fales, C. (2005). "Heaviness" in the perception of heavy metal guitar timbres: The match of perceptual and acoustic features over time. In P. D. Greene & T. Porcello (Eds.), *Wired For Sound: Engineering and Technologies in Sonic Cultures* (pp. 181–197). Middletown, CT: Wesleyan University Press.

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.

Burgoyne, J. A., & McAdams, S. (2008). A meta-analysis of timbre perception using nonlinear extensions to CLASCAL. In R. Kronland-Martinet, S. Ystad, & K. Jensen (Eds.), *Computer Music Modeling and Retrieval. Sense of Sounds* (pp. 181–202). Berlin, Germany: Springer Verlag Berlin Heidelberg.

Butler, M. J. (2006). *Unlocking the groove: Rhythm, meter, and musical design in electronic dance music*. Bloomington: Indiana University Press.

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

Cambouropoulos, E. (2009). How similar is similar? *Musicae Scientiae, 13*(1), 7–24.

Casey, M. A., 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), 668–696.

Collins, N. (2006). Towards a style-specific basis for computational beat tracking. In *Proceedings of the International Conference on Music Perception and Cognition (ICMPC)* (pp. 461–467).

Collins, N., Schedel, M., & Wilson, S. (2013). *Electronic music.* Cambridge, UK: Cambridge University Press.

Cui, B., Shen, J., Cong, G., Shen, H. T., & Yu, C. (2006). Exploring composite acoustic features for efficient music similarity query. In *Proceedings of the 14th ACM International Conference on Multimedia* (pp. 412–420). Retrieved from http://dl.acm.org/citation.cfm?id=1180725

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.

Dayal, G., & Ferrigno, E. (2013). Electronic Dance Music. *Grove Music Online*. Retrieved from http://oxfordindex.oup.com/view/10.1093/gmo/9781561592630.article.A2224259

de Haas, B., Rohrmeier, M., Veltkamp, R. C., & Wiering, F. (2009). Modeling harmonic similarity using a generative grammar of tonal harmony. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 549–554). Retrieved from http://ismir2009.ismir.net/proceedings/OS7-2.pdf

de Leon, F., & Martinez, K. (2012). Enhancing timbre model using MFCC and its time derivatives for music similarity estimation. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (pp. 2005–2009). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6333837&tag=1

Dixon, S., Gouyon, F., & Widmer, G. (2004). Towards characterisation of music via rhythmic patterns. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 509–516). Retrieved from http://ismir2004.ismir.net/proceedings/p093-page-509-paper165.pdf

Flexer, A. (2014). On inter-rater agreement in audio music similarity. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 245–250). Retrieved from http://www.terasoft.com.tw/conf/ismir2014/proceedings/T045_256_Paper.pdf

Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo (ICME)* (Vol. 1, pp. 452–455). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=869637

Foote, J., Cooper, M. L., & Nam, U. (2002). Audio retrieval by rhythmic similarity. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 265–266). Retrieved from http://ismir2002.ismir.net/proceedings/03-SP02-1.pdf

Foote, J. T. (2003). Media segmentation using self-similarity decomposition. In *Electronic imaging*. International Society for Optics and Photonics. Retrieved from http://spie.org/Publications/Proceedings/Paper/10.1117/12.476302

Gordon, J. W. (1987). The perceptual attack time of musical tones. *The Journal of the Acoustical Society of America, 82*(1), 88–105.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America, 61*(5), 1270–1277.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience, 5*(11), 887–892.

Grosche, P., Müller, M., & Serrà, J. (2012). Audio content-based music retrieval. *Multimodal Music Processing, 3*, 157–174.

Guastavino, C., Gómez, F., Toussaint, G., Marandola, F., & Gómez, E. (2009). Measuring similarity between flamenco rhythmic patterns. *Journal of New Music Research, 38*(2), 129–138.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning* (Unpublished master's thesis). The University of Waikato, Hamilton, New Zealand. Retrieved from http://www.cs.waikato.ac.nz/~mhall/thesis.pdf

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

Helmholtz, H. L. F., & Ellis, A. J. (1954). *On the sensations of tone as a physiological basis for the theory of music*. Cambridge, UK: Cambridge University Press.

Herre, A., Allamanche, E., & Hellmuth, O. (2001). Robust matching of audio signals using spectral flatness features. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (pp. 127–130). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=969559

Honingh, A., Panteli, M., Brockmeier, T., López Mejía, D. I., & Sadakata, M. (2015). Perception of timbre and rhythm similarity in electronic dance music. *Journal of New Music Research, 44*(4), 373–390.

Jayant, N. S., & Noll, P. (1984). *Digital coding of waveforms: Principles and applications to speech and video* (pp. 115–251). Englewood Cliffs, NJ: Prentice-Hall.

Jones, M. C., Downie, J. S., & Ehmann, A. F. (2007). Human similarity judgments: Implications for the design of formal evaluations. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 539–542). Retrieved from http://ismir2007.ismir.net/proceedings/ISMIR2007_p539_jones.pdf

Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 6, pp. 3089–3092). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=757494

Klapuri, A., Eronen, A. J., & Astola, J. T. (2006). Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech and Language Processing, 14*(1), 342–355.

Lamere, P. (2008). Social tagging and music information retrieval. *Journal of New Music Research, 37*(2), 101–114.

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: Design, validation and optimization. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 521–526). Retrieved from http://ismir2008.ismir.net/papers/ISMIR2008_145.pdf

Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects* (pp. 237–244). Retrieved from http://cms2.unige.ch/fapse/neuroemo/pdf/ArticleLartillot2007Bordeaux.pdf

Logan, B., & Salomon, A. (2001). *A content-based music similarity function* (Report No. CRL 2001/02). Cambridge Research Laboratory, Technical Report Series. Retrieved from http://www.hpl.hp.com/techreports/Compaq-DEC/CRL-2001-2.pdf

Longuet-Higgins, H. C., & Lee, C. S. (1984). The rhythmic interpretation of monophonic music. *Music Perception, 1*, 424–441.

Marsden, A. (2012). Interrogating melodic similarity: A definitive phenomenon or the product of interpretation? *Journal of New Music Research, 41*(4), 323–335.

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., Barrington, L., & Lanckriet, G. R. G. (2010). Learning similarity from collaborative filters. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 345–350). Retrieved from http://ismir2010.ismir.net/proceedings/ismir2010-59.pdf

McLeod, K. (2001). Genres, subgenres, sub-subgenres and more: Musical and social differentiation within electronic/dance music communities. *Journal of Popular Music Studies, 13*(1), 59–75.

Müllensiefen, D., & Frieler, K. (2004a). Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. *Computing in Musicology, 13*, 147–176.

Müllensiefen, D., & Frieler, K. (2004b). Optimizing measures of melodic similarity for the exploration of a large folk song database. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 274–280). Retrieved from http://ismir2004.ismir.net/proceedings/p052-page-274-paper178.pdf

Nash, C., & Blackwell, A. (2011). *Tracking virtuosity and flow in computer music*. Ann Arbor, MI: MPublishing, University of Michigan Library.

Novello, A., McKinney, M. M. F., & Kohlrausch, A. (2011). Perceptual evaluation of inter-song similarity in Western popular music. *Journal of New Music Research, 40*(1), 1–26.

Novello, A., van de Par, S., McKinney, M. M. F., & Kohlrausch, A. (2013). Algorithmic prediction of inter-song similarity in Western popular music. *Journal of New Music Research, 42*(1), 27–45.

Orpen, K. S., & Huron, D. (1992). Measurement of similarity in music: A quantitative approach for non-parametric representations. *Computers in Music Research, 4*, 1–44.

Pachet, F., & Aucouturier, J.-J. (2004). Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences, 1*(1), 1–13.

Painter, T., & Spanias, A. (2000). Perceptual coding of digital audio. *Proceedings of the IEEE, 88*(4), 451–515.

Pampalk, E. (2004). A Matlab toolbox to compute music similarity from audio. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 254–257). Retrieved from http://ismir2004.ismir.net/proceedings/p048-page-254-paper180.pdf

Pampalk, E., Flexer, A., & Widmer, G. (2005). Improvements of audio-based music similarity and genre classificaton. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 634–637). Retrieved from http://ismir2005.ismir.net/proceedings/1053.pdf

Panteli, M., Bogaards, N., & Honingh, A. (2014). Modeling rhythm similarity for electronic dance music. In *International Society for Music Information Retrieval Conference* (pp. 537–542). Retrieved from http://www.terasoft.com.tw/conf/ismir2014/proceedings/T097_268_Paper.pdf

Paulus, J., & Klapuri, A. (2002). Measuring the similarity of rhythmic patterns. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 150–156). Retrieved from http://ismir2002.ismir.net/proceedings/02-FP05-1.pdf

Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project* (Technical Report, IRCAM). Retrieved from http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf

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

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

Pohle, T., & Schnitzer, D. (2009). Submission to MIREX 2009 audio similarity task. In *Proceedings of the Music Information Retrieval Evaluation Exchange (MIREX)*. Retrieved from http://www.music-ir.org/mirex/abstracts/2009/PS.pdf

Pohle, T., Schnitzer, D., Schedl, M., Knees, P., & Widmer, G. (2009). On rhythm and general music similarity. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 525–530). Retrieved from http://ismir2009.ismir.net/proceedings/OS6-1.pdf

Reynolds, S. (2007). *Bring the noise: 20 years of writing about hip rock and hip hop*. London, UK: Faber & Faber.

Reynolds, S. (2012). *Energy flash: A journey through rave music and dance culture*. London, UK: Picador.

Rocha, B., Bogaards, N., & Honingh, A. (2013). Segmentation and timbre similarity in electronic dance music. In *Sound and Music Computing Conference* (pp. 754–761). Retrieved from http://smcnetwork.org/node/1830

Schedl, M., & Knees, P. (2009). Context-based music similarity estimation. In *Proceedings of the 3rd International Workshop on Learning Semantics of Audio Signals* (pp. 59–74). Retrieved from http://lsas2009.dke-research.de/proceedings.html

Schedl, M., & Knees, P. (2013). Personalization in multimodal music retrieval. In M. Detyniecki, A. Garcia-Serrano, & A. Nuernberger (Hg.), *Adaptive multimedia retrieval: Large-scale multimedia retrieval and evaluation: Proceedings of the 9th International Workhsop, AMR 2011* (pp. 58–71). New York: Springer.

Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America, 103*(1), 588–601.

Schnitzer, D., Flexer, A., & Widmer, G. (2012). A fast audio similarity retrieval method for millions of music tracks. *Multimedia Tools and Applications, 58*(1), 23–40.

Schroeder, M. R., Atal, B. S., & Hall, J. L. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America, 66*, 1647–1652.

Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. (2012). Unsupervised detection of music boundaries by time series structure features. In *Twenty-Sixth AAAI Conference on Artificial Intelligence* (pp. 1613–1619). Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/viewFile/4907/5309

Sethares, W. A. (1998). Consonance-based spectral mappings. *Computer Music Journal, 22*, 56–72.

Seyerlehner, K., Schedl, M., Knees, P., & Sonnleitner, R. (2011). A refined block-level feature set for classification, similarity and tag prediction. In *Proceedings of the Music Information Retrieval Evaluation Exchange (MIREX)*. Retrieved from http://www.music-ir.org/mirex/abstracts/2012/SSKS1.pdf

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4) 591–611.

Smith, L. M. (2010). Rhythmic similarity using metrical profile matching. In *International Computer Music Conference* (pp. 177–181). Retrieved from http://hdl.handle.net/2027/spo.bbp2372.2010.034

Stober, S., & Nürnberger, A. (2013). Adaptive music retrieval – a state of the art. *Multimedia Tools and Applications, 65*(3), 467–494.

Styles of house music. (2013). *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Styles_of_house_music

Sun, L., Ji, S., & Ye, J. (2013). *Multi-Label Dimensionality Reduction*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Terasawa, H., Slaney, M., & Berger, J. (2005). Perceptual distance in timbre space. In *Proceedings of the International Conference on Auditory Display (ICAD)* (pp. 61–68). Retrieved from http://hdl.handle.net/1853/50176

Thul, E., & Toussaint, G. T. (2008). Rhythm complexity measures: A comparison of mathematical models of human perception and performance. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 663–668). Retrieved from http://ismir2008.ismir.net/papers/ISMIR2008_125.pdf

Toiviainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huotilainen, M., & Näätänen, R. (1998). Timbre similarity: Convergence of neural, behavioral, and computational approaches. *Music Perception, 16*(2), 223–241.

Tzanetakis, G., & Cook, P. (2000). Audio information retrieval (AIR) tools. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*. Retrieved from http://ismir2000.ismir.net/papers/tzanetakis_paper.pdf

van Noorden, L., & Moelants, D. (1999). Resonance in the perception of musical pulse. *Journal of New Music Research, 28*(1), 43–66.

Vassilakis, P. N. (2001). *Perceptual and physical properties of amplitude fluctuation and their musical significance* (Doctoral dissertation). University of California, Los Angeles.

Vassilakis, P. N., & Kendall, R. A. (2010). Psychoacoustic and cognitive aspects of auditory roughness: Definitions, models, and applications. In *Proceedings of Human Vision and Electronic Imaging XV* (pp. 1–7). Retrieved from http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=776221

Vignoli, F., & Pauws, S. (2005). A music retrieval system based on user driven similarity and its evaluation. In *Proceedings of the International Conference of Music Information Retrieval (ISMIR)* (pp. 272–279). Retrieved from http://ismir2005.ismir.net/proceedings/1021.pdf

Volk, A., & van Kranenburg, P. (2012). Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae, 16*(3), 317–339.

Witek, M. A. G., Clarke, E. F., Wallentin, M., Kringelbach, M. L., & Vuust, P. (2014). Syncopation, body-movement and pleasure in groove music. *PLoS ONE, 9*(4), e94446.

Wolff, D., & Weyde, T. (2014). Learning music similarity from relative user ratings. *Information Retrieval, 17*(2), 109–136.

Yeston, M. (1976). *The stratification of musical rhythm*. New Haven, CT: Yale University Press.

Appendix: Dataset

Table A1. List of the commercially released tracks from which the 12-second segments have been extracted.

	Artist	Song	Start time of segment (s)
1	Aphex Twin	Cornish Acid	4.4
2	Cornelius	Breezin	40.7
3	The Prodigy	Firestarter	237.7
4	Burial	Loner	171.8
5	Amon Tobin	Get Your Snack On	128.0
6	Clark	Com Touch	207.0
7	Underworld	Crocodile	125.4
8	Afrojack	Die Hard	340.4
9	Massive Attack	Teardrop	0.6
10	Leftfield	Original	356.1
11	Daft Punk	Around the World	245.7
12	Deadmau5	Soma	263.2
13	Squarepusher	Fat Controller	170.7
14	Flying Lotus	Parisian Goldfish	20.3
15	Tiesto	Euphoria	323.7
16	UMEK	Efortil	294.3
17	Merzbow	Transformed Into Food	16.3
18	Autechre	Clipper	119.1
19	Ricardo Villalobos	Amazordum	16.3
20	Richie Hawtin	Minus-Orange 2	33.1

Note. The 20 segments lead to 190 pairs of segments. The start time of the segment is indicated in seconds from the beginning of the track.