Journal of New Music Research, 2015
Vol. 44, No. 4, 373–390, http://dx.doi.org/10.1080/09298215.2015.1107102

Routledge
Taylor & Francis Group

Perception of Timbre and Rhythm Similarity in Electronic Dance Music

Aline Honingh[1], Maria Panteli[1], Thomas Brockmeier[1], David Iñaki López Mejía[1] and Makiko Sadakata[1,2]

[1]*University of Amsterdam, The Netherlands;* [2]*Radboud University Nijmegen, The Netherlands*

(Received 5 June 2015; accepted 7 October 2015)

Abstract

Music similarity is known to be a multi-dimensional concept, depending among others on rhythm similarity and timbre similarity. The present study aims to investigate whether such sub-dimensions of similarity can be assessed independently and how they relate to general similarity. To this end, we performed a series of web-based perceptual experiments on timbre, rhythm and general similarity in electronic dance music. Participants were asked to rate similarities of music pairs on a 4-point Likert scale. The results indicated that the ratings in the three types of similarity did not completely overlap and that participants showed slight to fair agreement in their ratings in all conditions. Together, the results suggest that it is possible to assess sub-dimensions of similarities independently to some extent. Interestingly, general music similarity was not completely explained by the summation of timbre and rhythm similarity. Based on this, a novel hypothesis of how general music similarity follows from its contributing sub-similarities is proposed.

Keywords: music similarity, rhythm, timbre, inter-rater agreement, perception

1. Introduction

1.1 Music similarity

Music similarity plays an important role in everyday music listening. It is because of music similarity that music can be recognized to belong to a certain genre. Because of music similarity, a new verse/chorus can be identified and as such, we recognize and appreciate a structure in music. Computing music similarity is also extremely important in the digital era, because this could help us archiving various digitized music information. Furthermore, retrieving a piece of music from a database based on a similarity to a whistled or hummed pattern is also an application of music similarity that receives ongoing attention (Ghias, Logan, Chamberlin, & Smith, 1995; Nagavi & Bhajantri, 2014).

Music similarity has received a lot of attention in recent years. For example, a number of models on music similarity have been proposed (Novello, van de Par, McKinney, & Kohlrausch, 2013; Pampalk, 2004; Schnitzer, Flexer, & Widmer, 2012; Wolff & Weyde, 2014) next to perceptual experiments (Downie, Lee, Gruzd, & Jones, 2007; Jones, Downie, Ehmann, 2007; Novello, McKinney, & Kohlrausch, 2011). In general, these studies agree in showing that different dimensions contribute to music similarities, such as genre, tempo and timbre (e.g. Novello et al., 2011). It is rather surprising that most experiments have not included a definition of music similarity when studied with participants. For example, Novello et al. 2011 stated that 'We did explicitly not provide the participants with any definition of "similar"'. As such, the interpretation of this term is up to the participants, assuming that people have a clear idea of this term. However, this concept may not be as clear-cut as researchers hope for, as Pachet and Aucouturier (2004, p. 1) speak about 'music similarity in general (whatever this may mean)'. The current study aims at contributing to unravel this concept of music similarity.

Similarity can be defined as partial identity (Cambouropoulos 2009). Two musical pieces are similar if they share some of their properties. For example, minuets are similar in their structure while Jazz pieces can be similar

Correspondence: Aline Honingh Institute for Logic, Language and Computation, University of Amsterdam, Science Park 107, 1090 GE Amsterdam, The Netherlands. E-mail: A.K.Honingh@uva.nl
Present address: Maria Panteli, Queen Mary University of London, UK.

This article was originally published with errors. This version has been corrected. Please see Erratum (http://dx.doi.org/10.1080/09298215.2015.1126209).

in genre characteristics. As such, we know that a context influences the way we evaluate similarity of musical pieces (Berenzweig, Logan, Ellis, & Whitman, 2004; Cambouropoulos, 2009). Three types of context have been identified, namely music content related (e.g. rhythm, harmony, form), music context related (e.g. song lyrics, artist background), and user context related (e.g. mood, musical training) (Schedl & Knees, 2013). The current paper primarily focuses on the first, the music content related context. However, we are aware of influences of music context and user context and will look into this as well (Section 4.3).

It must be noted that perception of music content features is not completely independent. For example, since tempo influences rhythm (Handel, 1992; Honing, Deutsch, Honing, & Deutsch, 2013), rhythm can not unambiguously be defined without taking into account tempo. Furthermore, timbre is known to influence perception of rhythm: if there is a rhythmic pattern played by two instruments, it can be the timbre of the instruments that makes us perceive two streams instead of one (Bregman 1994). Indeed, some models include timbre features as one of their parameters for simulating rhythm perception (Lartillot, Eerola, Toiviainen, & Fornari, 2008; Panteli, Bogaards, & Honingh, 2014). Such interactions among acoustic features make it extremely challenging to understand the nature/concept of similarity judgements. Our approach is to explicitly evaluate perception of 'sub-similarities' (timbre, rhythm) and general similarity from the same piece of music.

Music sub-similarities are factors that influence and contribute to 'general' music similarity, but they can also have value on their own. For example, it has been said that 'music taste' is often correlated with timbre (Pachet & Aucouturier, 2004). Furthermore, rhythm similarity has been used for the classification of dance music (Chew, Volk, & Lee, 2005). Besides these practical applications of sub-similarities, research on sub-similarities may also explain aspects of music perception such as for example the finding that rhythm similarity can be described in terms of families of musical rhythms (Cao, Lotstein, & Johnson-Laird, 2014). Although computational models have been created for timbre and rhythm similarity (Aucouturier, Pachet, & Sandler, 2005; Guastavino, Gomez, Toussaint, Marandola, & Gómez, 2009; Pachet & Aucouturier, 2004; Paulus & Klapuri, 2002), explicit perceptual experiments on these sub-similarities have not been performed. Also, although it has been shown that general music similarity is interpreted consistently by listeners (Jones et al., 2007; Novello et al., 2011), this has never been studied in the case of timbre and rhythm similarity.

In our study we focused on rhythm and timbre similarity, and used the restricted domain electronic dance music (EDM) as experimental stimuli, because rhythm and timbre are the two most important musical dimensions in this genre (Butler 2006). In this way, we hope to find out how perception of sub-similarities contributes to that of general similarity.

1.2 Electronic dance music

Electronic Dance Music (EDM), or simply 'dance music' is a metagenre encompassing a heterogeneous group of musics made with computers and electronic instruments (McLeod 2001). Hundreds of subgenres[1] and hybrid genres have been created since the start of EDM in the 1980s (Dayal & Ferrigno, 2014). The continuous and rapid introduction of new subgenre names into electronic/dance music communities is equalled by no other type of music (McLeod 2001). Subgenre labels are however not always uniquely defined, sometimes triggering debate on musicological differences. Also, single labels have been used for more than one style, like for example 'garage' and 'hardcore' (Collins, Schedel, & Wilson, 2013).

In EDM, melodies, vocals and sound effects are layered over a steady beat. A notable feature of many EDM tracks is the use of sampling, a practice in which a discrete portion of sound is recorded and then inserted into a new piece of music (Dayal & Ferrigno, 2014). EDM tracks can be divided into the 'four on the floor' genres such as techno, house, trance, and the 'breakbeat-driven' genres such as jungle, drum 'n' bass, breaks etc. Four on the floor genres are typically characterized by a four-beat steady bass-drum pattern whereas breakbeat-driven exploit irregularity by emphasizing the metrically weak locations (Butler 2006).

In EDM, timbre and rhythm are promoted above melody and harmony (Reynolds 2008). Timbre stands out as a primary compositional parameter. It is seen as the criterion by which patterns may be differentiated most easily (Yeston 1976). Most of the timbre changes that occur in EDM involve an element either entering or leaving the mix.

Rhythm is typically based on the concept of a 'loop', a repeating pattern associated with a particular (often percussive) instrument or instruments (Butler, 2006; Collins et al., 2013). Structural changes in an EDM track typically consist of an evolution of timbre and rhythm as opposed to the usual verse-chorus division found in pop music (Butler 2006).

Similarity in EDM has been previously investigated by studying to what extent EDM subgenres have influenced each other (Collins 2012).

1.3 The present study

In this study, we measured three types of similarities (rhythm, timbre and general similarity) and studied interactions among them. Two measures of consistency check were applied, inter- and intra-rater agreement.

Many previous studies concentrated on the similarity of entire musical pieces (Pachet & Aucouturier, 2004; Schnitzer et al., 2012). However, this approach may be ambiguous when interpreting the results, because it is not known which part

[1]The terms 'subgenre' and 'style' will be used interchangeably here, so as to cite the used resources as closely as possible. With both terms we mean the stylistic attributes of the music without the cultural context.

of music was used for the similarity rating. Therefore, in this study, we will focus on the similarity of relatively short segments of music.

We expect that, in general, high similarity perception in any of the two sub-similarities would result in higher general similarity. However, it could be the case that the general similarity is a (weighted) sum of both sub-similarities, or that one sub-similarity overrule another. The perception of sub-similarities and its interaction with general music similarity is a whole new research area to explore, and we make the first step.

2. Methodology and experimental design

2.1 Experimental procedure

A series of web-based experiments have been carried out in this study, all addressing some form of music similarity. Two of the the experiments focused on timbre similarity, two on rhythm similarity and one on general music similarity. The initial goal of the experiments was to measure inter-rater agreement, the degree of agreement among participants, and intra-rater agreement, the degree of agreement among repeated ratings by a single participant (see Section 2.3 and Table 1).

All experiments started with an informed consent form with an overview of the purpose, the task, and the rights of the participants such as confidentiality. After accepting the conditions, the participants were taken to a set-up page where they were asked to adjust the volume to a comfortable level by providing an EDM track ('Kong' by Bonobo) that had the same loudness level as the experimental segments.

After this set-up stage, there was an explanation stage. Since participants might be unfamiliar with the term timbre, in the experiments on timbre, the participants read the following explanation: 'Timbre can be described as being the tone colour or sound quality'. For explaining the concept of rhythm and timbre similarity, we used synthesized example pairs (see Section 2.2.3). Participants were invited to listen to pairs of EDM segments while getting information about the similarity of the segments. Participants in the experiments on timbre similarity were presented with example pairs explaining the concept of timbre similarity; participants in the experiments on rhythm similarity were presented with example pairs explaining the concept of rhythm similarity. Similar pairs were accompanied by the text: 'We consider these segments to be similar in timbre/rhythm'. Non-similar pairs were accompanied by the text 'We consider these examples to be "not similar" in timbre/rhythm, meaning somewhere in the range between "somewhat similar" to "dissimilar"'. Participants could listen to the example pairs as many times as they needed in order to understand the concept of timbre/rhythm similarity.

Following previous experiments on (general) music similarity (Downie et al., 2007; Novello et al., 2011) the concept of general music similarity was not explained in the experiment, such as not to bias the participants towards certain dimensions but to assess the concept holistically.

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 similarity of the segments ('How similar are they?/How similar are they in their timbre/rhythm?'). A 4-point Likert scale was used for this end: (1) dissimilar, (2) somewhat dissimilar, (3) somewhat similar, (4) similar. An even scale was chosen such that participants could not choose a middle category but were forced to think about their choice a little harder when they were unsure. The Likert scale allowed us to calculate inter-rater agreement among many participants, to compare different ratings directly, and to infer similarity matrices from the ratings. Depending on the application scenario, other paradigms can be used for similarity analysis as well (Novello et al., 2011; Wolff & Weyde, 2014).

Participants were also asked for their confidence on the rating they selected on a 3-point scale (not confident, somewhat confident and confident). Participants who took part in an experiment on inter-rater agreement of timbre/rhythm similarity were asked to rate 60 pairs of segments (music material set 1, see Section 2.2.1). Participants in an experiment on intra-rater agreement were asked to rate 18 pairs of segments (music material set 2, see Section 2.2.2), and to repeat the experiment five times in a period of three weeks. The order in which pairs were represented was randomized every time.

After completing the ratings, participants found a questionnaire where they were asked about their age, gender, musical expertise in general and within the EDM spectrum. They were also asked to write down the strategies they used to give accurate ratings.

2.1.1 Pilot

A pilot experiment was conducted for timbre and rhythm similarity (Experiments 1 and 3), in which the above outlined procedure was followed. For this pilot, we invited colleague researchers and students. The average ratings of this were used to confirm the choice of the 20 segments that comprise music material set 1 (Section 2.2.1), and to select the segments music material set 2 (Section 2.2.2).

2.1.2 Participant recruitment

Participants for Experiments 1, 3 and 5 (see Table 1) were invited via several media. A link was given personally to some participants, posted on social networks, mailing lists, and internet forums both specialized and non-specialized in EDM to retrieve as much information as possible from different populations.

We intended for participant groups of the different experiments not to overlap. We invited participants to only one experiment (either 1, 3 or 5), and posted invitations for the different experiments to different websites. Furthermore, the length of experiments 1 and 3 was around 45 min, such that it is unlikely that participants completed more than one experiment.

Table 1. Overview of experiments done in this study. Inter-rater agreement (inter) and intra-rater agreement (intra) were measured.

	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
Similarity criteria	Timbre	Timbre	Rhythm	Rhythm	General
Music material set type	1	2	1	2	2
Analysis	inter	intra	inter	intra	inter

Participants for Experiments 2 and 4 (to measure intra-rater agreement) were invited personally. Since the goal of these experiments was to show that it is possible to provide consistent ratings, individuals with some musical experience were selected.

2.2 Stimuli

Two sets of stimuli were used.[2] Music material set 2 is a subset of music material set 1.

2.2.1 Music material set 1

The stimuli used in Experiments 1 and 3 were a set of 20 twelve-second segments from commercially released EDM tracks. All possible combinations of pairs (190 pairs in total) were used in the experiment. To select the 20 EDM segments we tried to satisfy the following criteria:

(1) The set is representative of EDM and balanced over the different subgenres.
(2) The tempo is within a range of 110–133 bpm.
(3) The set is balanced over different acoustical features in the dimensions of timbre and rhythm.
(4) The set includes pairs of music that fall in one of the following four categories:

- 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 first criterion was necessary in order to be able to make claims about the whole spectrum of EDM. However, as we saw in Section 1.2, hundreds of subgenres exist and subgenre labels are not always uniquely defined. Therefore, subgenre labels could not be used for the selection procedure. Instead we proceeded as follows. An EDM expert selected 120 segments of EDM tracks that were representative of EDM and balanced over the different subgenres, based on his own expertise. For these 120 tracks, we investigated the tempo and acoustical features. The selection of 20 segments was made afterwards.

The criterion on tempo was based on the finding that people tend to rate pieces that have a similar tempo as overall similar (Novello et al., 2011). Since tempo is an important factor in music similarity, it needs to be controlled and therefore we chose to select segments within a narrow tempo range.

Criterion 3 was important in order to have a variety of timbres and rhythms present in the stimuli. We scored each segment on different features. For timbre we used the features: Weak-Strong, Soft-Hard, Low Energy-High Energy, Colourless-Colourful, Cold-Warm, Dark-Bright, Acoustic-Synthetic, and Empty-Full (Alluri & Toiviainen, 2010). For rhythm we used the features: Symmetry, Complexity, Event density, Length of loop, and Syncopation (Panteli et al., 2014). The scores were perceptual ratings based on listening to the segments.

The 20 segments were chosen from the 120 tracks based on tempo and as wide as possible variety of acoustical features. Also, we checked whether the final set was still balanced over EDM subgenres by consulting the EDM expert.

Criterion 4 was important to be able to measure rhythm similarity independently from timbre similarity. If, for example, pairs are only present in the categories 'high timbre similarity–high rhythm similarity' and 'low timbre similarity–low rhythm similarity', we would not be able to distinguish timbre similarity from rhythm similarity. The pilot experiment confirmed that criterion 4 was sufficiently satisfied: the material set contained at least five pairs in each of the four categories.

Although the data have been carefully selected using the mentioned criteria, we have not labelled the stimuli with categories, as has been done in some other experimental studies. For example, Novello et al. 2011 used the categories 'genre', 'tempo: fast-slow' and 'timbre: primary instrument' to characterize their stimuli. The main reason for not doing this in our experiment is that it was not our primary goal to interpret the dimensions of rhythm and timbre (Section 4.4) according to these categories. Instead, our focus lies on the comparison of timbre, rhythm and general similarity, for which the above criteria and especially criterion 4 were most important.

The final set of 20 segments can be found in Table A1 in Appendix A. All sounds were converted to monaural Waveform Audio File Format (.wav) and MPEG-2 Audio Layer III (.mp3, 320 Kbps) for browser compatibility with HTML5. All had a fade-in and fade-out of 50 ms to prevent undesired pops and clicks and were all normalized (using Praat—Boersma & Weenink, 2015) within ±5 dB according to their mean loudness level.

In Experiments 1 and 3 every participant was asked to rate 60 randomly selected pairs from the 190 pairs in total. The

[2]Music copyright issues for purposes of the experiments were dealt with Buma/Stemra association, http://www.bumastemra.nl/en/

60 pairs that were rated by the *i*th participant did not overlap with the pairs of the $i^{\text{th}} + 1$ and $i^{\text{th}} + 2$ participant to ensure all 190 pairs were rated a similar number of times.

2.2.2 Music material set 2

In order to measure the intra-rater agreement in Experiment 2 and 4, as well as to measure general music similarity in Experiment 5, we created a smaller material set. This was primarily used for a small number of participants who were asked to rate the same set of pairs multiple times (Experiments 2 and 4). For the comparison between between general music similarity (Experiment 5) and timbre and rhythm similarity extra requirements to the material set were added. To be able to predict how timbre and rhythm similarity would contribute to general music similarity, it would be important that a considerable number of pairs exists in all four categories of condition 4 from Section 2.2.1.

We have selected 18 pairs from the 190 pairs from music material set 1, such that a balanced number of pairs would appear in the following categories:

- high rhythm similarity–high timbre similarity: 3 pairs
- high rhythm similarity–low timbre similarity: 3 pairs
- low rhythm similarity–high timbre similarity: 2 pairs
- low rhythm similarity–low timbre similarity: 3 pairs
- moderate rhythm similarity–moderate timbre similarity: 3 pairs
- moderate rhythm similarity–high timbre similarity: 1 pair
- moderate rhythm similarity–low timbre similarity: 1 pair
- high rhythm similarity–moderate timbre similarity: 1 pair
- low rhythm similarity– moderate timbre similarity: 1 pair

For making these groups, we used a pilot similarity rating study which used the material set 1. Similarity ratings ranged from 1 to 4 from high similarity (from 2.8 to 4), moderate similarity (from 2.2 to 2.8), to low similarity (from 1 to 2.2). Music material set 2 is shown in Table A2 in Appendix A.

2.2.3 Example pairs

For explaining the concept of rhythm and timbre similarity in the experiments, we used synthesized example pairs. Synthesized music, instead of real music clips, allowed us to control parameters for getting the desired sub-similarities. Therefore we could create pairs that could illustrate the difficult concept of a high similarity in one dimension together with a low similarity in another dimension.

Pairs with the following two similarity combinations were created: high rhythm similarity–low timbre similarity and low rhythm similarity–high timbre similarity. A pair with high rhythm and low timbre similarity consisted of the same rhythmic pattern but with different instruments, while a pair with low rhythm similarity and high timbre similarity consisted of the different rhythmic patterns but with the same instruments. In this way, the similar timbre example pairs used the same timbre and the similar rhythm example pairs used the same rhythm. We have chosen to represent similarity in this way because, while similarity is a subjective concept, the ultimate similarity, equality, is not. As mentioned earlier, participants may not be able to clearly separate sub-dimensions of music when judging sub-similarities. It might be hard to hear a high similarity in a particular dimension (e.g. rhythm) when the similarity in another dimension (e.g. timbre) is low. The equality would help the participants to understand that the dimension that they were asked to pay attention to, was similar after all.

2.2.4 Control pairs

By distributing the experiment on the Internet, chances of getting non-serious participants were increased. A first step of dealing with that was to not include any partially completed experiment runs. With this procedure we eliminated more than 200 incomplete responses. To further dismiss participants who have been answering questions randomly or non-focused, we have inserted a control pair of music segments, one instance at the start of the experiment, and one (the same pair) at the end. This pair was chosen to be very dissimilar, both in rhythm and timbre. The participants who did not give both pairs either the same rating or adjacent ratings on the scale of 1 to 4, were not included in the analyses. No participants were excluded based on this procedure.

2.3 Analysis methods for inter- and intra-rater agreement

The first analysis that was done after each experiment was to measure the degree of inter- and intra-rater agreement. We introduce here the methods that were used for this. Further analyses methods are reported in Section 4. A high degree of intra-rater agreement is a necessary property to reveal whether the perception of rhythm/timbre similarity is a stable phenomenon (Novello et al., 2011). A high degree of inter-rater agreement would support the possibility for the development of a global perceptual model of rhythm/timbre similarity (Novello et al., 2011).

Inter- and intra-rater agreement can be measured in different ways, depending on the nature of the data and the number of raters. Our data had an ordinal scale and a large number of raters, and thus most standard measures were not directly applicable. Therefore, we chose to use three different measures that can together give an impression of the rater agreement.

The first measure was Fleiss kappa (Fleiss 1971), following the approach by Jones et al. (2007). Fleiss kappa is defined as the degree of actual agreement above chance over the degree

Table 2. Interpretation of κ values according to Landis and Koch (1977).

κ	Interpretation
≤ 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

of attainable agreement above chance. This is described by the equation

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{1}$$

where

$$\bar{P}_e = \sum_{j=1}^{k} {p_j}^2, \tag{2}$$

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}, \tag{3}$$

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i, \tag{4}$$

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{5}$$

and where N represents the number of pairs (indexed by $i = 1 \ldots N$), n the number of ratings per pair, k the number of categories ($k = 4$ in our case, indexed by $j = 1 \ldots k$), and n_{ij} the number of ratings that assigned the ith pair to the jth category.

Due to the randomization in the experiments that used music material set 1 (i.e. 60 pairs were randomly taken from the total of 190 pairs), we did not obtain an equal number of ratings for all pairs, which is needed for calculation of kappa. Therefore we used the number of ratings from the pair that was rated least, and selected that number of ratings at random from each of the rest of the pairs. This allowed us to get a complete data set to be analysed. To have a more representative and reliable analysis of all the ratings obtained, we repeated this process 1000 times, and the mean kappa and p-value were computed. For the experiments using music material set 2, there was no missing data. Therefore, the results for kappa could be calculated on the whole set of ratings. Landis and Koch 1977 created a table for interpreting κ values (see Table 2).

The second measure of inter- and intra-rater agreement is based on correlation. We computed the correlation between the ratings of each participant with that of the 'average participant' of the experiment. The average participant ratings were simulated by averaging all ratings per pair. For the experiment using music material set 1, the average participant thus 'rated' all 190 pairs. For correlation with (real) participants who rated 60 pairs, only the ratings of the matching pairs were considered. The reported p-values are the average of all p-values.

The last and a more common measure to give an indication of the inter-rater agreement is the variance. The smaller the variance, the more consistent were the raters. The variance of random ratings, i.e. equal numbers of responses in each category, equals 1.25. The results of our variance measure should thus be interpreted with respect to the range [0, 1.25]. Both κ and the correlation range from –1 to 1. The differences in values between κ and the correlation is caused by many differences in calculation. One main difference is that correlation considers only relative position. For example, the ratings (1, 2, 1, 3) are considered perfectly correlated with (2, 3, 2, 4), while the ratings are all different and thus receive a low κ value.

3. Experiments and results

3.1 Experiment 1: Inter-rater agreement of timbre similarity

A large-scale online experiment was created to assess the perceived timbre similarity in EDM, and its inter-rater agreement. The experiment ran from March to June 2014.

3.1.1 Participants

A total of 62 complete responses was received from 31 female and 31 male participants between 19 and 65 years of age (mean = 27.1, SD = 9.1). Forty participants reported having received formal musical training, starting between the ages of 5 and 25 (mean = 10.6, SD = 4.8). The styles they received training in include classical, pop, rock and jazz. Twenty-one of all the participants stated that they work with music professionally, mainly as producer and audio engineer. Although 55 participants were familiar with EDM to variable degrees, they reported knowing less than 20% of the segments on average. Headphones or ear-buds were used by 54 participants, while nine reported using loudspeakers. One participant reported using both methods, and another participant used professional studio monitors.

3.1.2 Experimental procedure

The experimental procedure such as explained in Section 2.1 was followed. In the explanation phase, the participants were presented with example pairs of segments explaining the concept of timbre similarity. In the rating phase, each participant was asked to rate 60 randomly selected pairs from music material set 1. The task was, to rate the pairs on their timbre similarity ('Listen to the following audio clips. How similar are they in their timbre?'). Figure 1 presents an example screenshot of a trial.

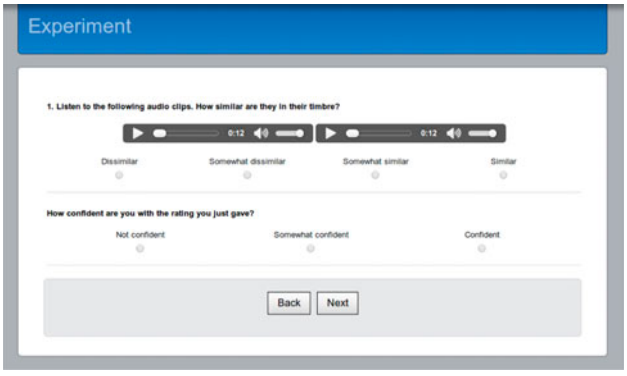

Fig. 1. Screenshot of fragment of the online experiment on timbre similarity.

3.1.3 Analysis and results

The results of inter-rater agreement measures introduced in Section 2.3 are summarized in the first column in Table 3. The kappa statistics indicated a slight agreement among participants, the correlation was 0.48 and the variance 0.91. The reason the agreement was rather low is not known. This could be due to the task difficulty: if the task is too difficult, a participant may have problems with performing it in a consistent manner. We will refer to this issue in Experiment 2. Another reason could be that sub-groups might exist that agree more amongst each other than the whole group does. In Section 4.3 we will go into detail about this option.

With respect to the strategies that the participants used for their ratings, few common methods were identified. First, participants tried to separate the sound into different instruments and to categorize the sounds in this way. The second method was to actually play the segments simultaneously. The third method was to compare the overall tone colour or feeling, and the fourth method that was identified was to classify the sounds, such as: warm, cool, distorted, acoustic. Surprisingly, rhythm and rhythmic elements were mentioned by some participants as well, while others stated that tempo and rhythm were difficult to ignore. We will further investigate listener strategies in Section 4.3.

3.2 Experiment 2: Intra-rater agreement of timbre similarity

To find out how reliably one can perform the similarity judgement of timbre, we asked three individuals to take part in the same experiment six times and assessed the intra-rater agreement. Experiment 2 used music material set 2 (Section 2.2.2).

3.2.1 Participants

Three participants took part in this experiment. The first participant was male, 29 years old, musically trained in classical music since the age of six. He played the flute, the piano and bass guitar. He reported being an EDM listener and researcher, and knew over 30% of the tracks used in the task. He used headphones to listen to the stimuli throughout the task.

The second participant was male, 22 years old, and trained in various styles of music since the age of eight. He played the piano and is somewhat familiar with EDM as a listener, knowing less than 10% of the tracks used in the experiment. He used ear-buds to listen to the music in the experiment.

The third participant was female, 26 years old, and trained in various musical styles since the age of eight. She played the flute and piano. She reported working with music professionally as a Deejay. She reported to be very familiar with EDM, and knew 10 to 20% of the tracks. She used headphones to listen to the music in the experiment.

3.2.2 Experimental procedure

The experimental procedure such as explained in Section 2.1 was followed. The participants were asked to rate the timbre similarity of 18 pairs of music material set 2. They only had to fill in the questionnaire once after the first session. A total of six experimental sessions took place in a period of three weeks. The order of the pairs was randomized every time.

3.2.3 Analysis and results

The measures of Fleiss kappa, correlation, and variance were calculated as described to measure the intra-rater agreement. The mean values for κ, correlation and variance are given in Table 3. The kappa statistics indicated a fair agreement among participants, whi ch is in proportion to correlation of 0.83 and a variance of 0.18.

With respect to the strategies, participant 1 stated that he 'tried to find similar instruments and compared their spectral components'. Participant 2 stated that he 'sometimes just followed his intuition' and sometimes tried 'to deconstruct the tracks to their various "channels" (like base drum, hi-hats, synthesizer) and compare their timbres'. Participant 3 stated that she was 'listening to instruments and sounds used'.

3.3 Experiment 3: Inter-rater agreement of rhythm similarity

Similar to Experiment 1, a large-scale online experiment was created to assess the perceived rhythm similarity in EDM and its inter-rater agreement. The experiment ran from March to June 2014.

3.3.1 Participants

A total of 57 complete responses was received from 13 female and 44 male participants between 17 and 52 (mean = 27.9, SD = 8.0). 31 participants reported having received formal musical training, starting between the ages of 4 and 20 (mean = 9.4, SD = 4.7). The styles they received training in include classical, jazz, pop, rock, and even electronical. Of

Table 3. Results of experiments done in this study. Experiments 1, 3 and 5 show the results of inter-rater agreement expressed in Fleiss kappa, inter-rater correlation and variance, while experiments 2 and 4 show the results of intra-rater agreement (the mean values are presented here). The *p*-values for Fleis kappa were all smaller than 0.001, and the *p*-values for the inter-rater correlation were all smaller than 0.05.

	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
Similarity criteria	Timbre	Timbre	Rhythm	Rhythm	General
Music data set type	1	2	1	2	2
Number of trials per participants	60	18	60	18	18
Number of participants	62	3	57	3	25
Number of repetitions per participant	1	6	1	6	1
(mean) kappa	0.06	0.30	0.16	0.38	0.16
(mean) inter-rater correlation	0.48	0.83	0.64	0.84	0.62
(mean) variance	0.91	0.18	0.78	0.18	0.65

all participants, 19 people stated that they work with music professionally, mainly as a musician. Although 51 participants were familiar with EDM to variable degrees, they reported knowing less than 20% of the segments on average. Headphones or ear-buds were used by 44 participants, while 16 reported using loudspeakers and some participants listened in multiple modalities.

3.3.2 Experimental procedure

The experimental procedure such as explained in Section 2.1 was followed. In the explanation phase, no explanation of rhythm similarity other than the example pairs were given. In the rating phase, each participant was asked to rate 60 randomly selected pairs from music material set 1. The task was, to rate the pairs on their rhythm similarity ('Listen to the following audio clips. How similar are they in their rhythm?').

3.3.3 Analysis and results

Values for Fleiss kappa, correlation and variance have been calculated, now applied to the data that resulted from the experiment on rhythm similarity. As can be seen from Table 3, participants agreed slightly on the task of rating rhythm similarity according to the Fleiss kappa measure, while the correlation was moderate, and the variance was 0.78. Comparing these results to the experiment on timbre similarity, we see that people agreed more in the dimension of rhythm than in that of timbre. We do not know yet whether this is because the task of rating rhythm similarity is easier than the task of rating timbre similarity, or perhaps because there are more strategies for rating timbre similarity.

The participants were asked what strategies they used in rating the pairs. A number of aspects were mentioned. Participants tried to compare the segments by reproducing, tapping along with or dancing to the rhythm. Some participants played the segments simultaneously. They also listened to different layers/streams of rhythm, and focused on rhythm aspects such as syncopation, time signature, periodicity, groove, swing and the rhythm pattern itself. Also tempo was mentioned as a factor. Some participants clearly stated that they tried to ignore the tempo and focus on rhythm regardless of tempo. Others mentioned that they used tempo as well. It was furthermore mentioned that it was difficult to ignore the timbre.

3.4 Experiment 4: Intra-rater agreement of rhythm similarity

This experiment investigated whether an individual participant could consistently rate 18 pairs of EDM segments with respect to rhythm similarity. Therefore, the intra-rater agreement was calculated on the basis of repeated rhythm similarity ratings by a participant. The participants were asked to rate the same pairs of segments six times. For this experiment, music material set 2 was used.

3.4.1 Participants

Three participants took part in this experiment. The first participant was a male, 25 years of age, trained as a musicologist, and received early-life musical training since the age of 7. He played the piano, guitar, drums and percussion and worked with music professionally. He reported being an EDM listener and knew between 20% and 30% of the tracks used in the task. He used headphones to listen to the stimuli throughout the task.

Participant 2 was a male, 36 years old, received musical training since the age of 8. He played the clarinet and reported to be unfamiliar with EDM, and knew less than 10% of the tracks. He used headphones to listen to the music.

Participant 3 was a female, 23 years old, received musical training since the age of 9. She played the violin and was very familiar with EDM as a listener, and knew between 20% and 30% of the tracks. She used headphones to listen to the music.

3.4.2 Experimental procedure

The participants were asked to rate the rhythm similarity of 18 pairs of music material set 2. They only had to fill in the questionnaire once after the first session. A total of six experimental sessions took place in a period of three weeks. The order of the pairs was randomized every time.

3.4.3 Analysis and results

The mean values for κ, correlation and variance are given in Table 3. The kappa statistics indicated a fair agreement among participants, which goes together with a correlation of 0.84 and a variance of 0.18. This shows that the task of rating rhythm similarity is, at least for these participants, not too difficult to perform.

With respect to the strategies, participant 1 stated that he was 'listening to complexity or regularity of different instruments, listening how beats are subdivided (more 8th notes or more triplets or more 16th notes), imagining how I would move along to it'. Participant 2 stated that he was listening to the 'bass rhythm, (non-) existence of regular pulse, and irregularities'. Participant 3 stated that she was 'tapping along with the rhythm, focussing on the bass or drums and trying not to listen to the melody-like instruments'.

3.5 Experiment 5: Inter-rater agreement of general music similarity

In this experiment, we had participants rating pairs of segments (music material set 2) with respect to general music similarity.

3.5.1 Participants

For this experiment, we have received complete responses from 25 participants (5 female, 20 male) between the ages of 18 and 78 years of age (mean = 31.95, SD = 14.35). A total of 14 participants reported having received formal musical education starting between the ages of 5 and 18 (mean = 9.5, SD = 3.11). The main instruments reported are guitar, piano and drums, with mentions of violin, recorder and voice. The styles they were educated in include classical, pop and rock, and also other genres were mentioned such as Hindustani classical and American folk. Eight participants work with music professionally, mainly as performers. Of all participants, 11 were very familiar and four were somewhat familiar with EDM. Eight people reported being producer and/or DJ within this genre. Only two were not familiar with EDM at all. The majority reported having listened to less than 20% of the segments before the experiment. Sixteen participants used headphones or ear-buds, 11 used loudspeakers, and two used professional studio monitors.

3.5.2 Experimental procedure

The experimental procedure such as explained in Section 2.1 was followed. Contrary to other experiments, no example screen was given, so the participants would not use any strategies specific to either timbre or rhythm. The task was, to rate the pairs on their similarity ('Listen to the following audio clips. How similar are they?'). All participants rated the 18 pairs from music material set 2. The order of pairs was randomized for every participant.

3.5.3 Analysis and results

We report the Fleiss kappa measure, correlation and variance of the data. As can be seen from Table 3, participants agreed slightly on the task of rating general similarity according to the Fleiss kappa measure, while the correlation was 0.62, and the variance was 0.65.

When asked for the participant's strategies that they used during the process of rating, rhythm aspects, such as time signature, drum pattern and regularity were reported the most. Other factors were, in order from most mentioned to least mentioned: instrumentation/timbre, tempo, genre/style, mood/feel, and melody.

4. Further analyses

In this section we will further analyse the data that resulted from the five experiments in this study.

4.1 Comparing experiments 1, 3, 5

First, we compare the results of Experiments 1, 3 and 5. For making fair comparisons, we have to first align the music materials. Since the material set 2 is a subset of set 1, it is possible to select pairs that were used in Experiment 5 from Experiments 1 and 3. Table 4 shows this comparison, namely, inter-rater agreement for timbre (Exp 1), rhythm (Exp 3) and general similarity (Exp 5), all using the same music material set 2. The measure of inter-rater correlation is not usable on this smaller set, as its results on timbre and rhythm similarity are not statistically significant. We see that the values for kappa and variance are roughly the same for all three types of similarity.

In the introduction we reasoned that general similarity contains sub-similarities like timbre and rhythm similarity. Therefore, as a concept, general similarity would be more complicated than timbre and rhythm similarity, and we might have expected that general similarity would have a lower inter-rater agreement than timbre and rhythm similarity. However, this was not supported in our findings.

For all types of similarity, the agreement on the confidence ratings was high, while the variance of the mean confidence values over all pairs was low (all confidence ratings were high on average). Therefore, there was no benefit in including the confidence ratings as weighting into the calculations for the inter-rater agreement (of similarity ratings). Instead, we have looked into the confidence ratings independently of the similarity ratings.

Analyses of the ratings of confidence showed that the confidence ratings of general similarity were significantly higher than the confidence ratings of timbre and rhythm similarity (Wilcoxon signed rank test $Z = 3.01$, $p < 0.01$ and $Z = 3.10$, $p < 0.001$ respectively). There was no significant difference between the confidence ratings of timbre and rhythm similarity ($Z = -1.07$, $p = 0.31$). This may mean that the

Table 4. Results of experiments 1, 3 and 5 on music material set 2.

	Exp 1	Exp 3	Exp 5
Similarity criteria	Timbre	Rhythm	General
kappa	0.16 ($p < 0.001$)	0.19 ($p < 0.001$)	0.16 ($p < 0.001$)
inter-rater correlation	0.64 ($p = 0.19$)	0.68 ($p = 0.16$)	0.62 ($p < 0.05$)
variance	0.77	0.73	0.65

Table 5. Average confidence ratings for pairs with varying similarity from the experiment on timbre similarity rhythm similarity, and general similarity.

average confidence rating of pairs rated as	timbre similarity (exp. 1)	rhythm similarity (exp. 3)	general similarity (exp. 5)
similar	2.79	2.81	2.91
somewhat similar	2.46	2.46	2.71
somewhat dissimilar	2.58	2.49	2.72
dissimilar	2.79	2.75	2.91

task of rating pairs on general similarity is on average easier than the task of rating pairs on timbre or rhythm similarity.

To compare confidence ratings with respect to the similarity categories, we calculated for each participant the average confidence for each similarity category (similar, somewhat similar, somewhat dissimilar, dissimilar). Taking these values for all participants together, we collected a group of confidence ratings for each similarity category. We compared the four groups of confidence ratings to every other group, using a Wilcoxon rank-sum test. The average ratings are shown in Table 5. The confidence values for ratings in the extreme similarity categories (similar, dissimilar) were significantly higher than the values in the middle categories (somewhat similar, somewhat dissimilar). The ratings in the extreme similarity categories did not differ significantly from each other, meaning that it was for people not necessarily easier to judge dissimilar pairs than similar pairs (or vice versa).

4.2 Interaction between rhythm, timbre, and general similarity

Before exploring how rhythm and timbre similarity interact with respect to general similarity, it is essential to compare the ratings of rhythm and timbre similarity to know whether people have rated them differently. To measure this, we use a Wilcoxon signed rank test, which can be used to compare related samples or repeated measurements. We averaged all timbre similarity ratings per pair and compared this to the averages of all rhythm similarity ratings per pair using the Wilcoxon signed rank test. The test indicated that the two experiment results were significantly different ($Z = -2.76$, $p < 0.01$). Comparing the rhythm and timbre similarity ratings that only apply to music material set 2, we did not find a significant difference ($Z = 0.588$, $p = 0.58$).

Figure 2 visualizes the mean ratings of the three experiments for all 18 pairs of music material set 2. We see that in Figure 2(a), the data is spread over the plane, indicating that pairs in all combinations of high/low timbre similarity and high/low rhythm similarity are present. We observe from Figures 2(b) and (c) that only the lower triangles of the plane are filled. This means that the general similarity of each pair is lower than or equal to its timbre or rhythm similarity.

This finding is confirmed if we look into the ratings in more detail. In Table 6 we listed some results per pair (using music material set 2). Here, the Wilcoxon rank-sum test was used to calculate whether the rhythm and timbre ratings were significantly different from the general similarity ratings for each pair, and to calculate whether the rhythm and timbre ratings were significantly different for each pair. Note that we did this test for each pair, using all ratings that were given by participants that apply to this pair (no averaging).

For nine pairs there was no significant difference between the ratings of timbre and rhythm similarity. The average similarity ratings are alike in this case, meaning that for these pairs the degree of rhythm and timbre similarity was similar (e.g. a pair may have exhibited both high timbre and high rhythm similarity).

For the other nine pairs, the timbre and rhythm similarity ratings were significantly different from each other. From these nine pairs, for eight pairs the general similarity ratings were significantly different from one of the dimensions (timbre or rhythm). This means that the ratings for general similarity were in agreement with either timbre or rhythm similarity. Zooming in to this agreement, we see that general similarity always seemed to be in agreement with (i.e. non-significantly different from) the dimension in which the particular pair was rated less similar. For one pair the general similarity was significantly different from both timbre and rhythm similarity. We assume that (an)other dimension(s) may have played a role here.

From this, we cannot say that general similarity in EDM is influenced more by either of the dimensions of timbre or

Table 6. For each of the 18 pairs of music material set 2, the following values are reported: (a) mean general similarity rating, (b) mean timbre similarity rating, (c) mean rhythm similarity rating (ratings between 1 (not similar) and 4 (similar)), all together with their standard deviation (SD), and average confidence value (conf.) for these ratings expressed in a number between 1 (low confidence) and 3 (high confidence); (d) the result on the test whether the rhythm/timbre results are significantly different (Y(yes)/N(no)—on the 0.05 level) from the general similarity results, and the result of the test whether the rhythm results are significantly different from the timbre results.

pair	(a) general sim (+SD)	conf.	(b) timbre sim (+SD)	conf.	(c) rhythm sim (+SD)	conf.	(d) diff. timbre-general?	diff. rhythm-general?	diff. timbre-rhythm?
1	1.80 (0.82)	2.84	2.50 (1.02)	2.79	2.35 (1.04)	2.70	Y ($U = 104.5$, $p = 0.031$)	N ($U = 173.5$, $p = 0.067$)	N ($U = 152.5$, $p = 0.662$)
2	2.08 (0.81)	2.64	2.62 (1.05)	2.52	2.17 (1.15)	2.61	N ($U = 256$, $p = 0.055$)	N ($U = 222.5$, $p = 0.959$)	N ($U = 322$, $p = 0.171$)
3	1.04 (0.20)	2.96	1.52 (0.77)	2.65	1.50 (0.83)	2.50	Y ($U = 251.5$, $p = 0.002$)	Y ($U = 183$, $p = 0.016$)	N ($U = 324$, $p = 0.759$)
4	2.40 (0.96)	2.80	3.20 (0.77)	2.67	2.42 (0.85)	2.36	Y ($U = 102$, $p = 0.013$)	N ($U = 171.5$, $p = 0.926$)	Y ($U = 154.5$, $p = 0.024$)
5	2.52 (0.92)	2.76	3.23 (0.86)	2.67	2.87 (0.99)	2.67	Y ($U = 216.5$, $p = 0.005$)	N ($U = 146.5$, $p = 0.230$)	N ($U = 273$, $p = 0.222$)
6	1.12 (0.44)	2.84	1.13 (0.35)	2.93	2.44 (1.04)	2.56	N ($U = 178.5$, $p = 0.648$)	Y ($U = 64.5$, $p < 0.001$)	Y ($U = 39$, $p < 0.001$)
7	1.72 (0.84)	2.84	2.97 (0.96)	2.70	1.82 (0.95)	2.65	Y ($U = 136.5$, $p < 0.001$)	N ($U = 202$, $p = 0.781$)	Y ($U = 403.5$, $p < 0.001$)
8	2.36 (1.04)	2.72	2.52 (0.89)	2.52	2.70 (0.66)	2.60	N ($U = 353.5$, $p = 0.551$)	N ($U = 207$, $p = 0.302$)	N ($U = 292$, $p = 0.709$)
9	1.28 (0.68)	2.92	1.69 (0.63)	2.62	1.17 (0.65)	2.83	Y ($U = 98.5$, $p = 0.017$)	N ($U = 308$, $p = 0.472$)	Y ($U = 225$, $p = 0.002$)
10	2.00 (1.04)	2.84	1.86 (1.10)	2.71	3.43 (0.81)	2.76	N ($U = 190.5$, $p = 0.636$)	Y ($U = 81.5$, $p < 0.001$)	Y ($U = 42.5$, $p < 0.001$)
11	1.84 (0.90)	2.68	2.88 (0.72)	2.81	3.55 (0.76)	2.90	Y ($U = 79.5$, $p < 0.001$)	Y ($U = 47$, $p < 0.001$)	Y ($U = 77.5$, $p = 0.005$)
12	1.68 (0.85)	2.96	3.07 (0.96)	2.73	3.20 (0.68)	2.60	Y ($U = 58$, $p < 0.001$)	Y ($U = 38.5$, $p < 0.001$)	N ($U = 108$, $p = 0.859$)
13	1.04 (0.20)	2.96	1.20 (0.48)	2.83	1.30 (0.80)	2.80	N ($U = 327$, $p = 0.138$)	N ($U = 221.5$, $p = 0.195$)	N ($U = 301$, $p = 0.988$)
14	2.20 (0.87)	2.76	2.00 (1.00)	2.76	3.06 (0.94)	2.44	N ($U = 240$, $p = 0.467$)	Y ($U = 111$, $p = 0.003$)	Y ($U = 69.5$, $p = 0.004$)
15	1.44 (0.71)	2.80	2.35 (0.95)	2.55	1.50 (0.89)	2.80	Y ($U = 183$, $p < 0.001$)	N ($U = 250.5$, $p = 1$)	Y ($U = 463$, $p = 0.002$)
16	3.32 (0.69)	2.88	3.46 (0.97)	2.85	3.72 (0.46)	2.78	N ($U = 131$, $p = 0.290$)	Y ($U = 154$, $p = 0.049$)	N ($U = 108.5$, $p = 0.687$)
17	2.00 (1.00)	2.76	2.33 (0.98)	2.60	2.53 (0.87)	2.65	N ($U = 148.5$, $p = 0.258$)	N ($U = 141.5$, $p = 0.059$)	N ($U = 111$, $p = 0.525$)
18	1.80 (0.87)	2.72	1.69 (0.95)	2.69	3.00 (0.75)	2.47	N ($U = 219.5$, $p = 0.580$)	Y ($U = 77$, $p < 0.001$)	Y ($W = 50$, $p < 0.001$)

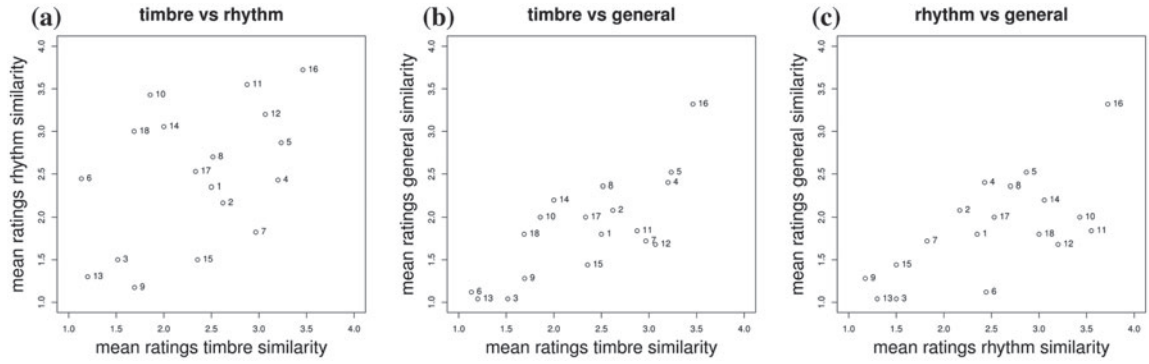

Fig. 2. Visualization of mean ratings for timbre, rhythm and general music similarity for the pairs of music material set 2.

rhythm similarity. Instead we see that general similarity takes either lower values than timbre and rhythm similarity or is non-significantly different from the dimension that was rated lowest.

We may hypothesize from this that general music similarity behaves like a logical AND-port where the output is only high (high general similarity) when all sub-similarities (like timbre and rhythm similarity) are high, and low when at least one of the contributing dimensions is low. Whether general similarity really behaves like this has to be determined by studying sub-similarities and interactions further in future research.

4.3 Listening strategies for timbre, rhythm and general similarity

From the results of the five experiments, we saw that there was some inter-rater agreement on the tasks of rating timbre similarity, rhythm similarity and general music similarity. However this agreement was rather modest. In this section, we will check if this relatively low agreement was due to user context (Schedl & Knees, 2013), such as musical training and familiarity with EDM. We checked the effect of user context on rating patterns.

4.3.1 Participant groups

We started by investigating the data of Experiments 1, 3 and 5 to check for differences in relation to musical background and experience with EDM. In the questionnaire that followed each experiment, the participants were asked to answer questions about, among others: (1) whether or not they had musical training, (2) how familiar they were with EDM, (3) how many of the musical fragments used in the experiment they heard before, and (4) whether or not they were working professionally with music. We hypothesized that these four attributes could have an influence on the similarity ratings.

For each type of similarity ratings we split the data into two groups using these four attributes: musical training versus no musical training; very familiar with EDM versus not very familiar with EDM; knowing more than 20% of the music in the experiment versus knowing less than 20%; working professionally with music versus not working professionally with music. For each group we calculated the variance per pair and the average variance. We compared two groups by comparing the variances using a Wilcoxon signed rank test. The results of these calculations can be found in Table 7. Although the variance can already be interpreted as a measure for inter-rater agreement, for comparison, we also reported on the inter-rater correlation for each group (as defined in Section 2.3). To counteract the problem of multiple comparisons, a Bonferroni correction was applied. Since for each dataset (timbre, rhythm and general similarity), four hypotheses of attributes (ways to split the data) were tested, we lower our significance level from $\alpha = 0.05$ to $\alpha = 0.05/4 = 0.0125$, and report on the group difference (yes/no) accordingly.

From Table 7 we can see that there is a significant difference between people who had musical training and people who did not, for the way they rated rhythm similarity. For rating timbre similarity and general similarity, musical training had no influence. Another difference was found between people who are very familiar with EDM (as listener or professionally) and people who are not, for the way they rated timbre similarity. For rating rhythm similarity and general similarity, familiarity with EDM had no influence. The fact whether or not participants were working professionally with music, and the percentage of music used in the experiment that they heard before, had no influence on any of the experiments.

4.3.2 Listening strategies

In addition to the idea that rating behaviour can be influenced by personal characteristics like musical training, it may also be the case that different listening strategies exist, such as for example having a focus on the low frequencies, or being particularly sensitive to vocals or syncopation. It is possible that such a listening strategy results from the participants primary instrument or musical training, but this connection is not considered here.

To get more insight into listener behaviour and particular listening strategies, we used cluster analysis. Cluster analysis is an unsupervised learning technique where a set of objects

Table 7. Overview of values for variance and inter-rater correlation of data from experiments on timbre similarity, rhythm similarity and general similarity, and the same values for a number of subgroups, for which is indicated whether there is a significant difference between the variances of the groups using a Wilcoxon signed rank test. Results for timbre and rhythm similarity are based on music material set 1, results for general similarity are based on music material set 2. Using a Bonferroni correction, group differences are reported to be significant when $p < 0.0125$.

	timbre similarity			rhythm similarity			general similarity		
	var.	cor.	p	var.	cor.	p	var.	cor.	p
whole music material set	0.91	0.48	< 0.01	0.78	0.64	< 0.01	0.65	0.62	< 0.05
musical training	0.92	0.50	< 0.01	0.81	0.66	< 0.01	0.68	0.66	< 0.05
no musical training	0.94	0.55	< 0.001	0.70	0.68	< 0.01	0.58	0.64	< 0.05
group difference?	no ($Z = 0.50$, $p = 0.62$)			yes ($Z = 2.87$, $p < 0.01$)			no ($Z = 1.24$, $p = 0.23$)		
very familiar with EDM	0.84	0.55	< 0.01	0.85	0.67	< 0.01	0.66	0.67	< 0.05
not very familiar with EDM	0.96	0.50	< 0.01	0.75	0.66	< 0.01	0.66	0.62	< 0.05
group difference?	yes ($Z = -2.96$, $p < 0.01$)			no ($Z = 1.56$, $p = 0.12$)			no ($Z = -0.20$, $p = 0.87$)		
know more than 20%	0.89	0.62	< 0.001	0.77	0.72	< 0.001	0.69	0.70	< 0.01
know less than 20%	0.91	0.49	< 0.01	0.78	0.65	< 0.01	0.64	0.59	< 0.05
group difference?	no ($Z = -1.21$, $p = 0.23$)			no ($Z = 2.14$, $p = 0.03$)			no ($Z = 0.41$, $p = 0.69$)		
working professionally with music	0.91	0.56	< 0.01	0.80	0.67	< 0.001	0.61	0.71	< 0.05
not working professionally with music	0.91	0.50	< 0.01	0.77	0.65	< 0.01	0.66	0.60	< 0.05
group difference?	no ($Z = -0.60$, $p = 0.55$)			no ($Z = 0.01$, $p = 0.99$)			no ($Z = -0.63$, $p = 0.54$)		

are grouped in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups. By choosing the participants as our objects, and letting a cluster algorithm decide on the similarity between the ratings of those participants, a dendrogram (hierarchical cluster) can result from our data. Participants that are close together in the tree, have given similar ratings to the pairs in the experiment.

In the experiment on general similarity, all participants rated all pairs of music material set 2. Therefore, we had a complete matrix of data for the cluster analysis. For the experiments on timbre and rhythm similarity, this was not the case. Participants rated only 60 of the total 190 pairs. Therefore, in the cluster analysis we had to deal with some missing data. We solved this by creating a distance measure that was based on the overlapping part of the ratings (i.e. the distance between two participants was calculated using only the ratings that applied to the same pairs).

We investigated clusters that were made with an average and complete linkage clustering method. The groups resulting from the cluster algorithm only serve as suggestions for possible listener groups. We used a Wilcoxon signed rank pair test to check whether the obtained groups were indeed significantly different, and tried to interpret the groups using the participant information from the experiment.

For general similarity, the cluster analysis resulted in two large groups plus a few listeners that did not belong to each of these groups (such that we do not consider these). These clusters were indeed found to be significantly different from each other. Turning to the demographical information that we had for the participants, we tried to interpret the clusters. Cluster 1, consisting of eight participants, contains seven musically trained people. With a Wilcoxon signed rank pair test we checked that indeed this group is not significantly different from the group of musically trained participants ($Z = 1.45$, $p = 0.16$). Group 2 included only four musically trained people out of a total of twelve participants, however, this group was significantly different from the non-musically trained participants ($Z = -3.55$, $p < 0.001$). A further interpretation by comparing the clusters to the listener groups considered before (see Table 7) and the group of timbre raters (experiment 1) and rhythm raters (experiment 3) could not be found.

For the participants who rated timbre similarity, two clear clusters (and only a few participants in a third cluster that we did not consider here) were found as well. Cluster 1, consisting of 31 participants, and cluster 2, consisting of 23 participants, were compared using a Wilcoxon signed rank test, and were shown to be significantly different from each other ($Z = -4.97$, $p < 0.001$). Both clusters were compared to the groups that were considered before (see Table 7), but the clusters could not be identified as one of these categories.

Looking into the strategies that the people in those clusters reported on, it was remarkable to see that cluster 1 reported on 'tempo' and 'beat', in addition to more timbre features. From this, it seems that cluster 1 was, besides listening to timbre features, also clearly taking into account temporal aspects of the music. Comparing the mean ratings of cluster 1 to the mean ratings of the participants who rated rhythm similarity, it was found that there was no significant difference (Wilcoxon signed rank test $Z = -0.038$, $p = 0.97$). This suggests that among the participants who rated *timbre* similarity, a number of them have used *rhythm* elements to rate the pairs. This could mean that it is difficult for some people to focus solely on timbre.

The dendrogram illustrating the hierarchical cluster of the participants who rated rhythm similarity, showed two main clusters as well. Cluster 1, consisting of 23 participants, and

cluster 2, consisting of 33 participants, were compared using a Wilcoxon signed rank test, and were shown to be significantly different from each other ($Z = 5.24$, $p < 0.001$). The clusters were compared to the clusters that were considered before (see Table 7), and it turned out that cluster 2 was not significantly different from the cluster of non-musically trained people who rated rhythm similarity. With respect to the strategies that the people reported on, both clusters reported on 'tapping along', 'time signature', 'tempo', 'rhythmic patterns', 'syncopation' and 'mixing'. Cluster 2 reported considerably more on 'rhythm' (in general) and 'movement/feeling', than cluster 1. This could be explained by the fact that cluster 2 was also identified to be non-significantly different from the cluster of non-musically trained people, since for people without musical training, a musical vocabulary to specify aspects of rhythm may be lacking.

4.4 Dimensions in rhythm and timbre similarity

In this section we try to identify specific aspects or dimensions that contributed to rhythm and timbre similarity by means of an exploratory approach. Since in Experiments 1 and 3 we have obtained ratings for all 190 pairs, we can fill a full 20×20 matrix that represents all distances (or inverse similarities) between all possible pairs of the 20 segments from music material set 1. The value at each element is the mean of the ratings for that particular pair. Two full distance matrices can be obtained, one for rhythm similarity and one for timbre similarity.

Like in the previous section, we will use clustering, now to represent clusters in the music segments instead of in the participants. Using clustering, we may be able to find out which segments of music belong together with respect to timbre and rhythm similarity, and as such discover what dimensions contribute to these kinds of similarity in EDM.

Here we present the dendrograms (trees that result from an hierarchical clustering method) that were made with the average linkage clustering method (also known as Unweighted Pair Group Method with Arithmetic Mean (UPGMA)).

Figure 3 shows the dendrogram for the segments of music from music material set 1 with respect to rhythm similarity where the average ratings of all participants have been used. As we saw in the previous section, for timbre similarity, a part of the participants might have used rhythm elements to rate the pairs on the basis of timbre similarity. Since these ratings would only distort the dendrogram here and make its interpretation increasingly difficult, we omit these participants here. The dendrogram in Figure 4 is made from the ratings of the cluster of participants that was indicated by cluster 2 in the previous section. The numbers in the dendrograms correspond to the segment numbers as we used them in the experiment (see Table A1 in Appendix A).

The dendrograms can be interpreted by listening to the segments of music and finding common characteristics among a cluster. Although the segments were rated according to certain timbre and rhythm categories in the process of selecting the

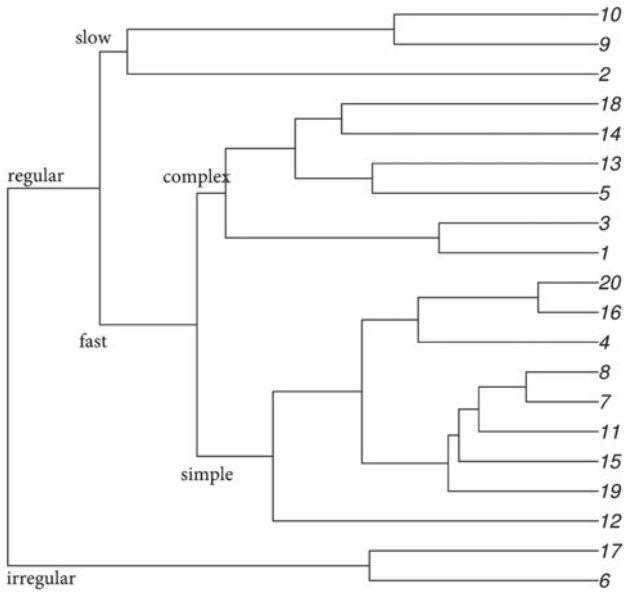

Fig. 3. Dendrogram illustrating clusters of music segments (music material set 1) with respect to rhythm similarity

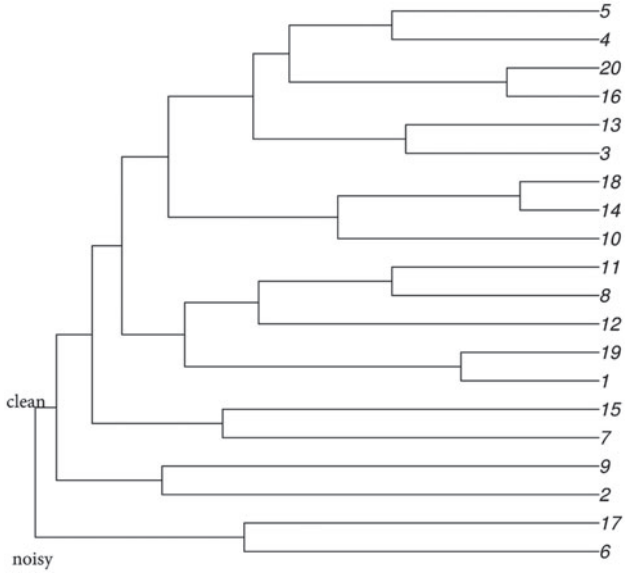

Fig. 4. Dendrogram illustrating clusters of music segments (music material set 1) with respect to timbre similarity

data (Section 2.2.1), these ratings were only used as a guide for the selection process and not for interpreting the dendrograms. Instead, the (groups of) segments of music were assessed holistically by listening, such as to only label groups that were clearly identifiable with a single label. The interpretation of clustering results is known to be a subjective process.

First looking at the dendrogram for rhythm similarity, we identified a first split indicating irregular rhythm (segments 6 and 17) versus regular/periodic rhythm. Then, there was a clear distinction in tempo, as segments 2, 9 and 10 were perceptually slower than the rest. Finally, there was a distinction in simple versus more complex (more syncopated) rhythm, where the four on the floor type segments were in one group

and the breakbeat segments in the other. A further interpretation of the subdivisions was difficult, and no convincing labels could be found.

The interpretation of the dendrogram for timbre similarity was less straightforward. One reason for this is that the inter-rater agreement for timbre similarity (and for the group of raters we use here) was lower than for rhythm similarity, such that different (unidentified) strategies may have played a role. Another reason may be that timbre similarity contained more dimensions than rhythm similarity, and that the music material set may have been too small to be able to observe these. The first three divisions in the dendrogram each separated two segments from the rest of the music. It is difficult to base an interpretation on only two segments, and hence it was not possible to unambiguously interpret this dendrogram. However, the first dimension division (segments 6 and 17 versus the rest) is most clear, and probably represents the division between noisy segments (6,17) and the cleaner/harmonic segments. To interpret more dimensions of timbre similarity in EDM, more music material seemed to be necessary.

5. Summary and discussion

In this study we did experiments to investigate timbre similarity, rhythm similarity and general music similarity. Inter-rater agreement was found for both timbre and rhythm similarity, the agreement for rhythm similarity being slightly higher than for timbre similarity. Intra-rater agreement values for each similarity showed that consistent similarity ratings were given. The intra-rater agreement was higher than the inter-rater agreement. Such a difference, namely that people agree more with themselves than with each other, suggests that personal rating behaviours do exist.

We addressed the question of whether it was possible to group these behaviours in Section 4.3. We found that two factors, namely musical training and familiarity with EDM had influenced the similarity ratings, but in different ways respectively. The results indicated that musical training was found to influence response patterns of rhythm similarity while familiarity with the EDM was found to influence timbre similarity. This seems to make sense, as analyses and judgements of rhythms may require more technical knowledge of music theory while that of timbre may require more experience with genres.

Comparison of general music similarity to timbre and rhythm similarity revealed that general similarity in EDM is not a weighted sum of its contributing sub-similarities. Instead we see that general similarity takes either lower values than timbre and rhythm similarity for a particular pair, or is non-significantly different from the dimension that was rated lowest. We may hypothesize from this that general music similarity behaves like a logical AND-port where the output is only high (high general similarity) when all sub-similarities (like timbre and rhythm similarity) are high, and low when at least one of the contributing dimensions is low. Whether general similarity really be haves like this has to be determined by studying sub-similarities and interactions further in future research.

We found that the task of rating pairs on general similarity tended to be easier than the task of rating pairs on timbre or rhythm similarity (people were more confident). Thus, although timbre and rhythm similarity may be more specifically defined than general similarity, participants were on average more confident to rate general similarity than timbre and rhythm similarity. Rating timbre and rhythm similarity did not differ significantly with regards to their difficulties as their confidence values did not differ.

From the clustering of rhythm similarity, we could clearly observe three dimensions, being regularity, tempo and complexity (foremost through syncopation). The interpretation of the dendrogram for timbre similarity was more difficult. Except for the dimensions of noisiness, which could be clearly observed, more data is needed to identify the consensus of the remaining dimensions.

In this study, we chose to use only 20 segments of music since this allowed us to have a reasonable number of people judge the same pairs of segments so as to be able to compute inter-rater agreement, which was our first priority. The 190 pairs that resulted from these 20 segments made up a full similarity matrix, which allowed us to look into the dimensions of rhythm and timbre similarity using clustering. However, since for timbre similarity we could not unambiguously interpret the dendrogram, one possible conclusion is that the music material set was not diverse and/or large enough to express these dimensions.

Knowing that tempo is an important dimension in music similarity we tried to exclude tempo as a dimension, by choosing the musical segments within a narrow tempo range (see Section 2.2). However, the cluster analysis (Section 4.4) suggested that the little tempo variation that was left within the data, was used (and reported on) by (part of) the participants. Thus tempo appeared to be a significant cue when judging the music similarity.

The results of the analysis on listener strategies (Section 4.3.2) indicated that, from the participants who were invited to rate timbre similarity, almost half of the participants have used rhythmic elements to rate the pairs. This raises the question of participants' understanding and perception of timbre. One possibility to account for this is that they did not understand our explanation and examples of timbre similarity. Another possibility is that it is so hard to ignore the rhythmic features when judging timbral aspect of music. Yet another possibility is that they truly believe that the rhythmic features that they used were included in the concept of timbre. In any case, this is an interesting observation and further study is necessary.

Besides the contribution of this study for the understanding of concepts of (timbre and rhythm) similarity, the dataset of participants ratings could also be used for testing computational models of timbre and rhythm similarity. Data (participants ratings) are available upon request.

6. Concluding remarks

We argued that music similarity should be treated as a multi-dimensional concept, and that the investigation of sub-similarities like timbre and rhythm similarity and its interaction with general music similarity could be a fruitful start to disentangle the complex nature of music similarity and the perception hereof. From the observed interactions, we hypothesize that general music similarity may behave like a logical AND-port, where the output gets high only when all sub-similarities are high, and gets lower when at least one of the contributing dimensions is low. This is a testable hypothesis, and further research will hopefully offer useful insights.

Acknowledgements

The authors wish to thank Niels Bogaards for helping to create the music material sets. Furthermore, we would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the manuscript.

Funding

The first author is supported by the Netherlands Organization for Scientific research (NWO-VENI grant 639.021.126). The second author was supported by a grant from the Centre for Digital Humanities Amsterdam.

References

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

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), 1–8.

Berenzweig, A., Logan, B., Ellis, D.P., & Whitman, B. (2004). A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal, 28*(2), 63–76.

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer [Computer program] (Version 5.4.16). Retrieved 16 August 2015 from http://www.praat.org/

Bregman, A.S. (1994). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

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

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

Cao, E., Lotstein, M., & Johnson-Laird, P.N. (2014). Similarity and families of musical rhythms. *Music Perception: An Interdisciplinary Journal, 31*(5), 444–469.

Chew, E., Volk, A., & Lee, C.Y. (2005). Dance music classification using inner metric analysis. In *The next wave in computing, optimization, and decision technologies* (pp. 355–370). Berlin: Springer.

Collins, N. (2012). Influence in early electronic dance music: An audio content analysis investigation. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)* (pp. 1–6). Canada: International Society for Music Information Retrieval.

Collins, N., Schedel, M., & Wilson, S. (2013). Electronic dance music. *Electronic music* (Chapter 8, pp. 102–119, Cambridge Introductions to Music Series). Cambridge: Cambridge University Press.

Dayal, G., & Ferrigno, E. (2014). *Electronic dance music*. Oxford: Oxford University Press.

Downie, J.S., Lee, J.H., Gruzd, A.A., & Jones, M.C. (2007). Toward an understanding of similarity judgments for music digital library evaluation. In *Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries* (pp. 307–308). New York: ACM.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378.

Ghias, A., Logan, J., Chamberlin, D., & Smith, B.C. (1995). Query by humming: musical information retrieval in an audio database. In *Proceedings of the third ACM international conference on multimedia* (pp. 231–236). New York: ACM.

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

Handel, S. (1992). The differentiation of rhythmic structure. *Perception & Psychophysics, 52*(5), 497–507.

Honing, H. (2013). The structure and interpretation of rhythm in music. In D. Deutsch (Ed.), *Psychology of music* (pp. 369–404). New York: Academic Press.

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 Society for Music Information Retrieval (ISMIR)* (pp. 539–542). Vienna: Austrian Computer Society (OCG).

Landis, J.R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: Design, validation and optimization. *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. Canada: International Society for Music Information Retrieval. (pp. 521–526)

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.

Nagavi, T.C., & Bhajantri, N.U. (2014). Progressive filtering approach for query by humming system through empirical mode decomposition and multiresolution histograms. *Journal of Intelligent Systems, 24*(2), 265–275.

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

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.

Pampalk, E. (2004). A Matlab toolbox to compute music similarity from audio. *Proceedings of the International Society for Music Information Retrieval (ISMIR)* (4pp). Barcelona: Universitat Pompeu Fabr.

Panteli, M., Bogaards, N., & Honingh, A. (2014). Modeling rhythm similarity for electronic dance music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)* (pp. 537–542). Canada: International Society for Music Information Retrieval.

Paulus, J., & Klapuri, A. (2002). Measuring the similarity of rhythmic patterns. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)* (7pp). Paris: IRCAM – Centre Pompidou.

Reynolds, S. (2008). *Energy flash: A journey through rave music and dance culture* (2nd ed.). London: Picador.

Schedl, M., & Knees, P. (2013). Personalization in multimodal music retrieval. In *Adaptive multimedia retrieval. Large-scale multimedia retrieval and evaluation* (pp. 58–71). Berlin: Springer.

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.

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 A. Data

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

	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

Table A2. List of the 18 pairs that form music material set 2.

pair number	track 1	track 2
1	Aphex Twin – Cornish Acid	Afrojack – Die Hard
2	Aphex Twin – Cornish Acid	Autechre – Clipper
3	Cornelius – Breezin	Burial – Loner
4	The Prodigy – Firestarter	Amon Tobin – Get Your Snack On
5	The Prodigy – Firestarter	Squarepusher – Fat Controller
6	The Prodigy – Firestarter	Ricardo Villalobos – Amazordum
7	Burial – Loner	Massive Attack – Teardrop
8	Burial – Loner	Deadmau5 – Soma
9	Clark – Com Touch	Tiesto – Euphoria
10	Underworld – Crocodile	Ricardo Villalobos – Amazordum
11	Afrojack – Die Hard	Tiesto – Euphoria
12	Daft Punk – Around the World	Tiesto – Euphoria
13	Daft Punk – Around the World	Merzbow – Transformed Into Food
14	Tiesto – Euphoria	UMEK – Efortil
15	Tiesto – Euphoria	Autechre – Clipper
16	UMEK – Efortil	Richie Hawtin – Minus-Orange 2
17	Autechre – Clipper	Ricardo Villalobos – Amazordum
18	Ricardo Villalobos – Amazordum	Richie Hawtin – Minus-Orange 2