A Study on the Metrical Structure of Music with Similarity Experiments

Stephan Hübler, Rüdiger Hoffmann

Laboratory of Acoustics and Speech Communication, Technische Universität Dresden

Stephan.Huebler@tu-dresden.de

Abstract

This study explains the metrical hierarchical structure of music that deals with onsets of musical events and the distances between them. The metrical structure is essential for the temporal organization of notes and the perception of rhythm. The paper shows the suitability of inter onset interval histograms to cover major metrical characteristics such as dominant metrical levels, regularity and onset density of a musical piece. Furthermore, experiments on automatic music similarity retrieval demonstrate the usefulness of the inter onset interval histogram for real world applications. The user chooses one song of a music collection for which he wants to get similar songs. The retrieval system sorts all other songs of the music collection according to the similarity of metrical properties to the given song. The results support the suitability of the histogram feature to retrieve similar songs within a rhythmic music collection.

Index Terms: music, rhythm, metrical structure, inter onset interval histogram, music similarity retrieval

1. Introduction

Rhythm describes the temporal phenomena of music and speech [1]. Even though the metrical organization of music is quite obvious in comparison to speech, the speech domain may still benefit from the study of music signals, which is the subject of this paper. London makes a clear distinction between rhythm and metre [2]. The metrical structure forms the internal framework to attend the perception and cognition of rhythm by humans. Rhythm in contrast "involves patterns of durations that are phenomenally present in music", writes London. The focus of this study is the metrical hierarchical structure of music, which deals with possible points in time for musical onsets and their distances. A common way to capture temporal properties of music is to build an inter onset interval histogram, that shows the distances between onsets [3, 4, 5]. This paper provides a detailed interpretation of histograms from songs of different music genres. We show the suitability to represent the metrical hierarchical structure of a musical piece with inter onset interval histograms.

Another aim of the study is to show the usefulness of the histograms for real world applications like an automatic similarity recommendation system. The goal of such a system is to find similar music titles from a database to a song of the database selected by the user called the seed song [6, 4]. The R60 database with manual onset labels and the ballroom dance set with automatic onset labels are used for experiments [7, 8]. Both databases involve music genres with strong rhythmic characteristics and consist of recordings with drums. A song is considered similar to the seed song when it is from the same music genre. The results of the similarity experiments support the use of inter onset interval histograms as feature for metric similarity retrieval.

2. Metrical structure of music

The organization of notes in music, especially in Western popular music, is not arbitrary. Their organization within a hierarchical metrical system is one major aspect of rhythm in music [9, 10]. The alternation of strong and weak beats form regular pulse rates that lead to internal expectations about the timing of upcoming musical events. Several regular pulse rates coexist that form the metrical hierarchical structure. The most obvious pulse rates from top to bottom are: bar, tactus and tick as shown in figure 1. The bar boundary pulse rate categorises the time signature. The tactus pulse rate is the time interval that the music listener would tap along with a finger or foot and most of the time this is equivalent to the quarter note. The tick describes the shortest time interval that is present in the musical piece.

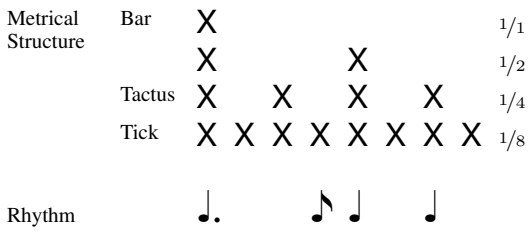


Figure 1: Metrical hierarchical system of music.

The note length of neighbouring hierarchical levels are connected with divisor respectively factor two for binary music. It is obvious that the metrical levels in figure 1 form just the basic metrical structure. Other regular pulse rates, for example from polyphonic rhythmic structures, are possible. The tempo of the piece of music determines the duration between the points of time defined by the metrical levels. It is specified by the number of beats per minute (bpm) on the tactus level.

3. Inter onset interval histogram

An onset marks the starting time of a musical event [11]. This could be a stroke on the drum, a pluck on the guitar or the starting point of a singing voice. An onset is characterized with a sudden change of the spectral energy. The placement of the labels by humans is considered as the most accurate possibility to receive the labels. Only a very limited number of annotated databases exists [11, 12, 7]. Since manual labeling is a time consuming process, there are many algorithms for automatic onset detection [11].

The inter onset interval histogram shows the occurrence of time spans between onsets and describes the periodicities of a musical piece. In literature, the histogram is starting point for music similarity calculations [4], dance style recognition [13], music genre recognition [5] and tick calculation [3]. The creation of the histograms is realized according to [3]. It uses not only the time span of consecutive onsets but all onsets in a window of 5.12 s. This time span falls within the scope of the upper

limit of rhythm perception between 5 – 6 s and leads to a deeper understanding of the metrical structure [2]. The two possibilities of windowing are:

1. `sliding (sl)`: Window moves to every onset.
2. `overlap (ol)`: Window moves with 50% overlap.

The histogram with windows of 50% overlap has a bias to shorter distances. This must not be a disadvantage because faster metrical levels are more important for the human listener, since they are processed in the short term memory (<3 s). The final histogram is the mean of all windows normalized to the maximum. Additionally it is smoothed with a Gaussian window. This way an onset interval influences the neighbouring time spans, which reflects the timing variation of the musicians with respect to the metrical structure.

4. Representing the metrical structure with inter onset interval histograms

The inter onset interval histogram reveals information about dominant metrical levels, the regularity of music, and the density of onsets. They are explained in detail in the next subsections with the help of six representative songs from the R60 database (see figure 3). The section ends with some comments on the tempo dependency of the histograms.

4.1. Dominant metrical levels

Figure 4.1 shows a tango musical piece with annotations of some metrical levels. The highest peak equates the tempo of the song, which is 124 beats per minute (bpm). Those represent the quarter notes. The tick is the first peak with 248 bpm. The listener of the song notices that this shortest note duration is equal to the 8th note. Other metrical levels are half notes at 62 bpm and the bar boundaries at 31 bpm. The automatic assignment of peaks to a specific metrical level remains a challenge. For example the highest peak does not necessarily belong to the tactus level.

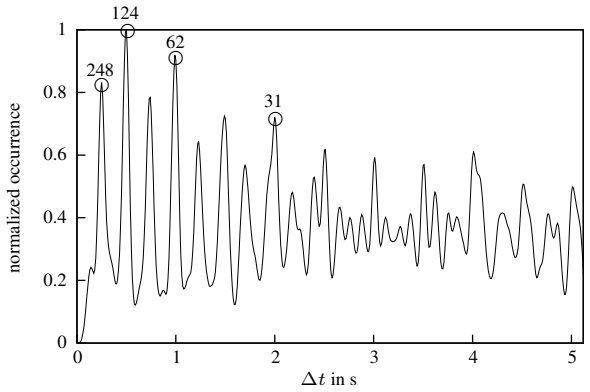


Figure 2: Inter onset interval histogram with interpretation of the most dominant metrical levels in beats per minute (bpm) of a Tango excerpt with `sl` windowing. (`Tan032`).

4.2. Regularity

Regularity describes the link among different metrical levels with a constant factor. Pieces with a high regularity stick very closely to this concept. The cha cha cha piece `Cha110` in figure 3(a) shows a high regularity. Another evidence for this characteristic is the big space between the upper and lower envelope of the histogram. In contrast, the hip hop piece `Hip006` in figure 3(e) has a low regularity. The space between the upper and lower envelope is small. Hip hop consists of many voice onsets. Those are hard to label and the rapper varies the rhythm towards the rhythmical structure and therefore nearly every onset distance is possible. It is worth to mention that dominant metrical levels are still visible as high peaks.

4.3. Onset density

The onset density relates to the distance of neighbouring peaks in the histogram. The two influencing factors are the tick and the tempo. The tick grid has a special meaning because it specifies most of the note onsets points in time [14]. The electronic music piece `Ele003` in figure 3(d) has a tick that corresponds to a 16th note at a tempo with 128 bpm. In comparison, the cha cha cha piece `Cha110` in figure 3(a) has a tick belonging to the duration of an 8th note at a comparable tempo of 124 bpm. Every second beat is missing in the cha cha cha piece compared to the electronic one.

The second factor is the tempo. The same piece of music at different tempi leads to a shorter distance of the peaks for the faster version. This has to be kept in mind when it comes to similarity retrieval. Histograms within the same tempo region are more likely to be similar. This must not be a disadvantage since the same tempo of musical piece is an important factor for similarity. This issue is explained in detail in [4].

5. Similarity experiments

This section examines the suitability of the inter onset interval histogram for a real world application: inter song similarity retrieval. For this application the algorithm should deliver similar songs to a seed song for a finite music collection [6]. Therefore a similarity measurement is needed. The similarity retrieved from the mean inter onset interval histograms μ with B bins of two songs s_1 and s_2 is calculated with the squared euclidean distance

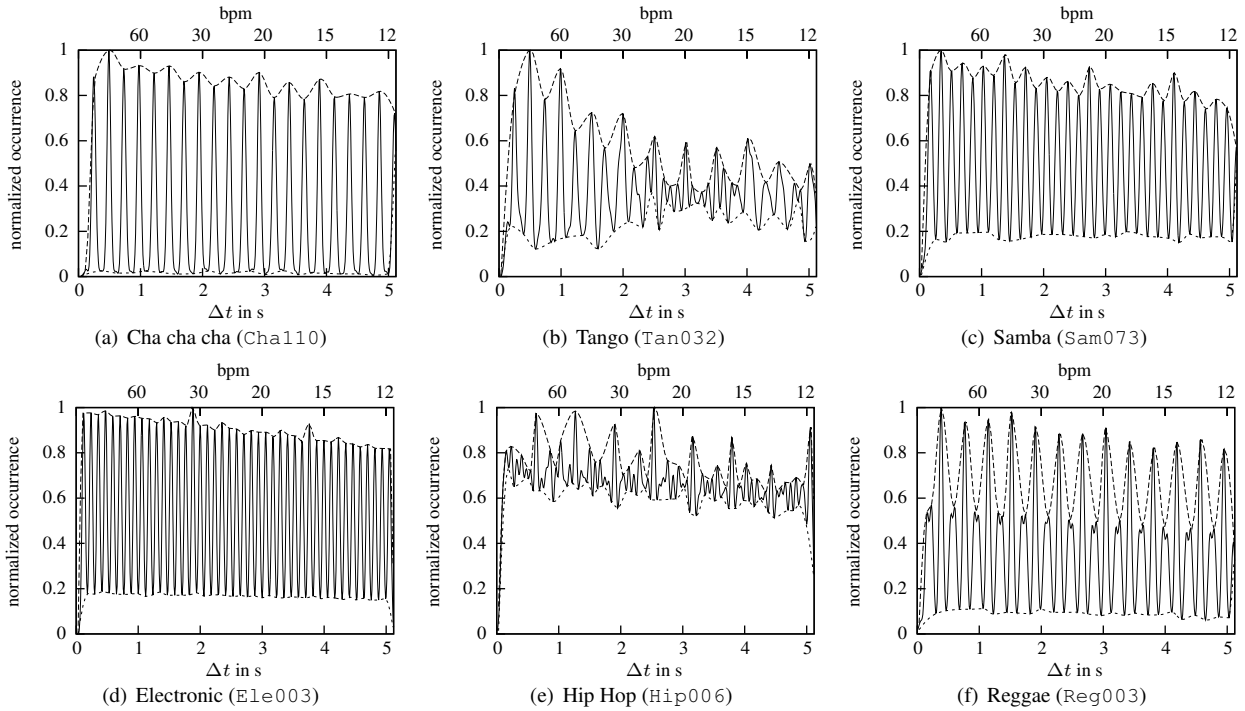
$$d(s_1, s_2) = \sum_{b=1}^{B} (\mu_{s_1}(b) - \mu_{s_2}(b))^2 \,. \qquad (1)$$

To get the similarity list for a seed song, the distances from the seed song to all other songs in the database are calculated. The application sorts all songs according to the distance to the seed song. The assumption is that songs with a small distance are more similar than those with a large distance. The system could "recommend" the first N files of the list as similar songs to the user.

5.1. Evaluation

Similarity judgement of songs is subjective. Listener `A` may find two songs similar – listener `B` might highly disagree. They might use different characteristics of the music for similarity judgement. Those could be timbre, instruments, rhythm, lyrics or artist to name a few. A way to judge similarity of songs is a listing test [15]. Since listening tests are time consuming other automatic evaluation methods are needed. A frequently used method considers the music genre of the songs. The genre describes various characteristics of a musical piece that makes it possible to categorize it into a style like rock, pop, reggae, tango, etc. . A common way to evaluate similarity systems is to view songs from the same genre as similar.

Since the inter onset interval histogram covers the metrical structure of music it will focus on the temporal characteristics of the music, leaving timbre features aside. Therefore it makes sense to use music collections consisting of genres where



(a) Cha cha cha (`Cha110`) (b) Tango (`Tan032`) (c) Samba (`Sam073`)

(d) Electronic (`Ele003`) (e) Hip Hop (`Hip006`) (f) Reggae (`Reg003`)

Figure 3: Inter onset interval histograms of six different genres from R60 database with upper and lower envelope and `sl` windowing.

rhythm has a major impact. The experiments are carried out on the R60 database [7] and the Ballroom dance set BRD [8]. All music titles of those databases involve drums as major rhythmic instrument. Manually labeled onsets exists for the R60 database. An automatic onset detection algorithm generates the onsets for the ballroom dance set.

A database consists of M music titles. Every title of the database is taken as a seed song to calculate the associated similarity list. The set $\mathbb{L}_N(s_m)$ covers N titles of the similarity list with smallest distance d to the seed song s_m. Those songs are the N similarity recommendations for song s_m. $\mathbb{G}(s_m)$ is the set of songs from the database with the same genre as the seed song s_m. The correctness C stands for the rate of correct recommendations within the N most similar songs retrieved for every song of the database.

$$C = \frac{\sum_{m=1}^{M} |\mathbb{L}_N(s_m) \cap \mathbb{G}(s_m)|}{N \cdot M} \,. \qquad (2)$$

5.2. Results manual labels – R60 database

The R60 database consists of six genres cha cha cha, tango, samba, electronic, hip hop and reggae with ten songs each and a duration of 30 s per song. The onset labels for the histogram calculation were placed by human listeners. The correctness for different N best songs of the similarity shows table 1.

The correctness decreases with the number of recommendations N. This is not surprising, since there are just nine titles for every seed song of the same genre. The correctness is always clearly above the baseline correctness of 0.15 when assigning titles randomly to the similarity list. The overlapping windowing `ol` leads to the better results compared to the sliding window `sl`. This supports the thesis that faster metrical levels are more important for similarity measurements. The confusion matrix in figure 4 shows the performance for every genre with three

	N				
	1	2	3	4	5
`sl`	0.78	0.73	0.69	0.65	0.60
`ol`	0.83	0.82	0.80	0.74	0.69

Table 1: Correctness of N recommendations to every song of the R60 database. The histogram is generated with two different types of window (see section 3).

entries per similarity list and `ol` windowing.

For tango the recommendations are perfect with every song belonging to the seed song genre tango. The algorithm works well for cha cha cha and hip hop. The hip hop genre is special. Most of the confusions of the genres samba, electronic and reggae are with hip hop. The example histogram for hip hop in figure 3(e) reveals less distinct peaks. In combination with the euclidean distance all other histograms are more likely to be similar to this kind of histogram. Other distances or additional derived features from the histogram might overcome this problem.

5.3. Results automatic labels – ballroom dance set

No manual onset labels exist in real world scenarios for music similarity recommendations. In addition the databases are way bigger than the R60 database. Experiments on the ballroom dance set should give an impression of the scalability and the effect of erroneous automatic onset detection [8]. The BRD has 689 song excerpts with a length of 30 s from eight different ballroom dances cha cha cha, jive, quickstep, rumba, samba, tango, viennese waltz and slow waltz.

Several automatic onset detection algorithms are available [11]. The mpeg7 feature audio spectrum envelope is starting point for the detection function used. Onsets are detected at

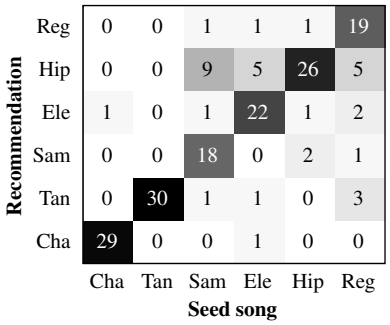


Figure 4: Confusion matrix for the R60 dataset with $M = 60$ songs and $N = 3$ recommendations per song (`ol` windowing). It shows the genre of the seed song and the genre of every recommended song.

points in time with sudden spectral energy changes. For more details on the algorithm see [7]. The correctness with $N = 5$ recommendations per song and `ol` windowing is 0.73 .

The correctness increases to 0.78 when the inter onset interval for the histogram is weighted with the lower energy of the two involved onsets. Onsets are not equally "important". The idea is that important onsets, like the beginning of a bar, have a high energy. This is caused by multiple instruments playing at this point in time. The distance between important onsets is more relevant than between an important or unimportant one, since they reflect the most important metrical levels. Therefore the inter onset interval is weighted with the minimum energy of the involved onsets. The detection function provides the energy value.

The results are comparable to the R60 dataset result, despite the utilized automatic onset detection algorithm. In fact, the correctness for R60 with $N = 3$ and `ol` windowing drops slightly from 0.80 to 0.73 correctness with automatic onset detection. The effect might not be as clear for BRD since every genre comprises more songs. A scalability seems possible, but since real world databases are dramatically bigger than BRD, the proof remains an open issue. The results for the BRD vary for different genres. From the worst for rumba with 0.52 correctness to 0.88 for cha cha cha. Automatic similarity recommendation based on the inter onset interval histogram is possible for rhythmic music. One has to keep in mind that the involved genres are quite consistent according to the tempo. This makes the usage of the inter onset interval histogram feature and the euclidean distance appropriate [4].

6. Conclusion

The onsets of notes in a musical piece form a hierarchical metrical structure. After an explanation of this structure we showed that the inter onset interval histogram is able to represent characteristics of this structure, such as dominant metrical levels, regularity of a musical piece and the density of onsets. The usefulness of the feature shows the application of music similarity retrieval. The task is to recommend similar songs of a rhythmic music collection for a given seed song. The R60 database with manual labels leads to a correctness of 0.80 and for the larger ballroom dance set with automatic onset detection the correctness is 0.78 . Improvements might consider the tempo dependency of the histograms and other distance measurements. Further research could try to adapt the ideas to speech signals were the rhythmical structure is not as obvious. Brown draws parallels from speech to music related to the organization of speech in metre [16]. The inter onset interval histogram captures the metrical properties of music and enables the design of a music similarity retrieval system for rhythmic music with drums.

7. References

[1] S. Hübler and R. Hoffmann, "Comparing the rhythmical characteristics of speech and music - theoretical and practical issues," in *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, ser. Lecture Notes in Computer Science (LNCS), A. Esposito, A. Hussain, M. F. Zanuy, and A. Nijholt, Eds. Berlin: Springer-Verlag, 2011, vol. 6456, pp. 376–386.

[2] J. London, *Hearing in Time – Psychological Aspects of Musical Meter*. Oxford University Press, 2004.

[3] F. Gouyon, P. Herrera, and P. Cano, "Pulse-dependent analyses of percussive music," in *AES, 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002

[4] M. Gruhne, C. Dittmar, and D. Gaertner, "Improving rhythmic similarity computation by beat histogramm transformations," in *ISMIR, 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, 2009, pp. 177–182

[5] T. Lidy, A. Rauber, A. Pertusa, and J. M. Inesta, "Combining audio and symbolic descriptors for music classification from audio," in *MIREX, 3rd Music Information Retrieval Evaluation eXchange*, 2007

[6] T. Magno and C. Sable, "A comparison of signal-based music recommendation to genre labels, collaborative filtering, musicological analysis, human recommendation and random baseline," in *ISMIR, 9th International Society for Music Information Retrieval Conference*, Philadelphia, USA, Sep 2008, pp. 161–166

[7] S. Hübler and R. Hoffman, "Evaluation of onset detection algorithms in popular polyphonic music on a large scale database," in *Proc. of the Audio Engineering Society 130th Convention*, London, UK, 2011.

[8] A. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," in *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 5, 2006, pp. 1832–1844.

[9] F. Lehrdal and R. Jackendoff, *The Generative Theory of Tonal Music*. MIT Press, 1983.

[10] F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Computer Music Journal*, vol. 29, no. 1, pp. 34–35, Mar 2005

[11] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in musical signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, Sep 2005

[12] P. Leveau, L. Daudet, and G. Richard, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *ISMIR, 5th International Society for Music Information Retrieval Conference*, Barcelona, Spain, 2004, pp. 72–75

[13] F. Gouyon and S. Dixon, "Dance music classification: A tempo-based approach," in *ISMIR, 5th International Society for Music Information Retrieval Conference*, Barcelona, Spain, 2004, pp. 501–504

[14] J. Bilmes, "Timing is of the essence: Percptual and computational techniques for representing, learning, and reproducing expressive timing in percussive music," Master's thesis, MIT, 1993.

[15] A. Novello, M. M. McKinney, and A. Kohlrausch, "Perceptual evaluation of inter-song similarity in western popular music," *Journal of New Music Research*, vol. 40, no. 1, pp. 1–26, 2011

[16] S. Brown and W. Kyle, "Speech is heterometric: the changing rhythm of speech," in *Speech Prosody, 5th International Conference*, Chicago, USA, 2010.