Extraction and Visualization of the Repetitive Structure of Music in Acoustic Data
–Misual Project–

Bartłomiej Puzoń
AGH University of Science and Technology, Cracow, Poland, and
NTT Communication Science Laboratories
bpuzon@o2.pl

Naoko Kosugi
NTT Communication Science Laboratories, 3-1, Wakamiya, Morinosato, Atsugi, Kanagawa, 243-0198, Japan
kosugi.naoko@lab.ntt.co.jp

ABSTRACT
This paper proposes a method to identify and visualize repetitive structures in music in PCM format using a similarity matrix. The Hough transform is used on the matrix to identify repetitions in a tune. Identified structures are visualized on an image called misual. Misual is a colored cylinder of a varying diameter, where colors represent repetitions and the diameter represents volume changes. The performance of the identification and visualization is evaluated with well-known classical tunes. Misual shows not only the dominant repetitions but also repeated patterns which can not be easily recognized while listening to a tune or reading its score. By revealing these patterns, misual gives a new point of view for musical structure analysis and at the same time gives people a new way to appreciate music.

Categories and Subject Descriptors
H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing

Keywords
music, repetition identification, visualization

1. INTRODUCTION
The musical structure, from short pattern repetitions to large musical forms (such as sonatas), is an important, complex characteristic of music. Repetitions are among the most significant elements of the structure. Their number, frequency and length may determine whether a piece will be easy to remember or hard to follow. They may invoke feelings of comfort and familiarity. By analyzing repetitions we may discover why a composition influences us the way it does. We may also imagine how an unknown composition would influence us if heard. Revealing repetitiveness structure opens for us a new world of music appreciation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

iiWAS2011 5-7 December, 2011, Ho Chi Minh City, Vietnam
Copyright 2011 ACM 978-1-4503-0784-0/11/12 ...$10.00.

This paper is a part of the Misual [10, 8] project. In the project, we are trying to overcome the natural limitations related to the traditional way of becoming familiar with a music piece - by listening. Due to the time required to listen to a piece, the amount of examples one can consider while, for example, searching for a composition having a certain property (e.g., mood) from online music databases is limited. Moreover, it is difficult to compare compositions with each other as our recollection about the particulars of a tune tends to fade as we listen to others. Misual aims to overcome these limitations by creating visualizations that allow the user to imagine the features of a musical piece and deduce their influence on a listener - all without the need to listen to the piece beforehand.

In this paper, we propose a way to automatically identify repeating fragments within a composition and visualize them in an intuitive way. We look for similar audio frames based on their Mel-Frequency Cepstrum Coefficients (MFCC). We encode distances between MFCC vectors assigned to each frame in a similarity matrix. Then, we convert the matrix into an image where similar audio fragments manifest themselves as lines. We propose the Local Minimum Filter algorithm to isolate the lines on the image and then perform the Hough transform to extract them. We use the MFCC distance information to further improve the extraction result. We treat the extracted lines as a set of constraints describing a music piece and we alter this set (adding lines or changing their length) so that the constraints do not contradict each other. The altered line set is then processed to extract repeat patterns information, including the structure of nesting. The set of identified repetitions is filtered to allow intuitive presentation and visualized in the misual framework, revealing repeated patterns as short as 1 second. This paper is organized as follows. Section 2 describes the research done in this area. Sections 3 and 4 explain in detail our method and its implementation. Sections 5 and 6 analyze the performance of the algorithm and the visualization on a set of examples in Table 1 and propose further research on the subject. Finally, Section 7 concludes the paper.

2. RELATED WORK
The aspect of similarity between audio segments is present in research mostly in two common contexts. First, similarities between different tunes can be analyzed in order to identify different instances of the same composition - Cheng uses it to allow audio queries into a music database [13]. Second, the similarities are identified within a tune to look for repetitive structures in music. The repetition data is
then used for music thumbnailing [1, 4, 5] or structural visualization [9].

The majority of similarity-based applications follow a similar framework. First, the audio data is divided into frames. Then, a measure for comparing frames with each other (usually MFCC) is used to create a self-similarity matrix with frame-to-frame distances [6]. The matrix contents may be additionally processed afterwards (automatic clusterization [1], distance thresholding [10]).

To identify repeat components, the similarity matrix is either scanned as it is [5, 10] or converted into an image and processed with image processing algorithms, like the Hough transform [1, 13]. The identified repeat components are then stored into a database for further reference [13] or visualized in 2D [5] or even 3D [9]. Previous papers written as part of the Misual project further analyze the visualizations of structure, volume, speed and note ratio content of music from the psychological point of view [10, 8].

3. ALGORITHM

The structure identification method we propose is based on the frame-to-frame similarity expressed by the distance between MFCC vectors assigned to audio frames. The distances are collected into a similarity matrix, which is further converted into an image with darker pixels indicating smaller distance (Fig. 1a). If two segments of music are similar, the frame-to-frame distance between respective frames should be relatively small and the resulting pixels relatively dark. Thus, a repetition of a pattern manifests itself on the image as a group of darker pixels, forming a line parallel to the diagonal. We convert the grayscale image into a binary one (Figs. 1b, 1c) and use the Hough transform backed up by the original distance data to retrieve a line set representing repetitions in the music. Then it is adjusted so that it represents valid repetitions and converted into repetition structure data which is in turn visualized in the misual framework.

3.1 Audio frame similarity measure

An MFCC set is computed for each frame of the input audio [10]. Afterwards, the Euclidean distance \( d_{ij} \) between each pair \((i,j)\) of frames is computed. The smaller the value \( d_{ij} \) is, the more similar the audio frames \(i\) and \(j\) are and the more similarly they should sound.

3.2 Graphical representation of similarity

The distances obtained in subsection 3.1 are shown on a 2D chart in Fig. 1a. Axes \( x \) and \( y \) represent the number of audio frame and the color of the pixel indicates the distance. We call the obtained image a grayscale similarity image. Let \( A \) be a sequence of audio frames \( A = (A_1, A_2, ... A_n) \). Any pair of subsequences \( (A_x, A_{x+1}, ..., A_{x+n}) \) and \( (A_y, A_{y+1}, ..., A_{y+n}) \) is represented by a line parallel to the diagonal in a similarity matrix. The similarity between the corresponding audio frames \( (A_x, A_y) \), \( (A_{x+1}, A_{y+1}) \), ..., is indicated by the color of the pixels along this line. Thus, darker diagonal lines on the similarity matrix indicate more similar audio subsequences. The degree of similarity is proportional to the color (two sequences that are exactly the same are represented by a completely black line). In order to use established methods of line extraction, we convert the grayscale image into a binary (black and white) one so that only the darkest lines (representing audio segment pairs with highest similarity) remain.

3.3 Obtaining the binary similarity image

We obtain the binary similarity image by setting each pixel color value to either black or white depending on the MFCC vector distance represented by it. The threshold used to judge what color a pixel should be is called a distance threshold \( T_h \). On the binary image only the pixels representing \( d_{ij} < T_h \) are retained. Figs. 1b and 1c show the result of the conversion with two different values of \( T_h \). The lower the threshold, the less information is preserved on the binary image. As the optimal \( T_h \) value is different for each tune, we propose a way to approximate it automatically. Our approximation is based on the distribution of distance values present in a tune. A typical histogram of the distribution is shown on Fig. 2. To find \( T_h \), we use the characteristics of the slope of the histogram and another threshold \( T_h \). \( T_1 \) values represent the relative height of the histogram. Pilot research showed that the best values of \( T_h \) are obtained with \( T_1 \) of 0.625, which ensures at least 90% recall for all the tunes we use. Thus, on the slope having a value of 0.625 of the histogram maximum should be taken to obtain the \( T_h \) for a given tune (as illustrated on Fig. 2).

3.4 Cleaning up the binary similarity image

In order to use the Hough transform to automatically find lines on the image, the rectangle-shaped clusters of black pixels (see the top images of Fig. 3) have to be removed from the image so that they do not generate the Hough transform’s false positives. To remove these clusters and isolate lines on the image, we propose an algorithm we call the Local Minimum Filter (LMF). The LMF algorithm removes those pixels of the binary image that could not have been a part of any distinguishable line on the grayscale image. The idea of the algorithm bases on the fact that the lines appear on the grayscale image because the pixels forming those lines are darker than the pixels around them (Fig. 1a). Specifically, if we draw a short line that is perpendicular to and crosses a given dark line, the pixel...
lying on the original line is the darkest point on the newly drawn one. In terms of underlying data, for a given (short enough) neighborhood \( k \), the pixels lying on darker lines on a similarity image hold the equation:

\[
d_{ij} = \min_{-N \leq k \leq N} d_{(i+k)(j-k)}
\]

where \( d_{ij} \) is the distance represented by a given pixel \((i,j)\). We check eq. 1 for every pixel of the binary image and if the relationship is not hold, the pixel is removed. As a result, the rugged, thick lines visible on the binary similarity image become thin as shown on Fig. 3a. Also, the lines hidden entirely in dark rectangular clusters, if any, become revealed (compare Fig. 3b and 3c).

![Figure 3](image-url)  
(a) (b) (c)

3.5 Extracting and merging lines

When the image is preprocessed, the Hough transform can be used to extract the lines. First, we have to pick parameters of the quantization that converts the \( xy \) space into a quantized distance-angle \( \rho \theta \) grid. The distance \( \rho \) dimension should contain the range corresponding to the whole image \((0 < \rho < L * \sqrt{2}, \text{ where } L \text{ is the image size})\). The resolution of \( \rho \) should be high in order to get better spatial localization of lines (i.e., better temporal localization of the repeat components). The possible angle \( \theta \) values should be limited so that only the lines parallel to the diagonal are considered \((45 - \alpha < \theta \leq 45 + \alpha, \text{ where } \alpha \text{ is a small value})\).

Occasionally (e.g., if applied to the LMF algorithm’s output similar to the one in Fig. 3b), the Hough transform may result in a large number of short line segments, usually shorter than the repetitions they are to represent. We adjust these segments using data from the original binary image. We analyze every pair of the lines \((l_1,l_2)\) and the line \( l \) created by merging them. The line pair \((l_1,l_2)\) is replaced by the new line \( l \) if the following two conditions hold:

1. The angle of \( l \) belongs to a given range \((45-\beta_1, 45+\beta_2)\).
2. At least \( TH_c \times \text{length}(l) \) pixels of the line are black on the binary image for a given threshold \( TH_c \).

If a given line \( l_1 \) can be merged with more than one \( l_2 \), the longest resulting \( l \) is chosen.

3.6 Internal transitivity of a line set

The set of lines obtained in the previous step represents repeated fragments in audio data. Before extracting the structural information about repetitions, we make the set consistent by introducing what we call internal transitivity. Let \( l = ((x_1,y_1),(y_2,z_2)) \) denote a line between points \((x_1,y_1) \) and \((x_2,y_2)\). The presence of such \( l \) on the image indicates that the sequence of audio frames \((A_{x_1},...,A_{x_2})\) is similar to the sequence \((A_{y_1},...,A_{y_2})\). We say that a set is internally transitive if the constraints it represents are transitive, i.e., the existence of \( l_1 = ((x_1,x_2),(y_1,y_2)) \) and \( l_2 = ((x_2,x_3),(y_2,y_3)) \) implies the existence of \( l_3 = ((y_1,y_2),(z_1,z_2)) \). Introducing internal transitivity results in adding lines to the set or increasing the length of selected lines. Figs. 4a and 4b show typical examples of line set alterations. The elements added to insure internal transitivity are shown in red.

![Figure 4](image-url)  
(a) (b) (c) (d) (e)

3.7 Identifying structure

Having an internally transitive set of lines ensures that, if a repeated segment starts in a given audio frame \( x_1 \), the line \( x = x_1 \) drawn on the image connects all the occurrences of the repeat pattern (Fig. 4c). We use this property to retrieve the structure of the repetitions. First, we remove duplicate information from the set, deleting all the lines that do not correspond to the first occurrence of the repeated pattern (Fig. 4d). If a component is repeated in succession, part of a line has to be removed as shown on Fig. 4e. The repeat components are then extracted from the line set as follows:

1. Scan the image along the horizontal time axis \((x)\).
2. If there is at least one line that begins in a given time frame \( x_1 \), create a new component candidate (a temporary structure used later to generate proper repeat components). Assign all lines that begin at \( x_1 \) to the core line set of the component candidate. Assign all lines that already existed at \( x_1 \) to the additional line set of the component candidate.
3. For each line that ends in a given time frame \( x_1 \), get a list of the component candidates that have this line assigned to any of the sets. For each component candidate found:

   (a) Create a new repeat component based on it.
   (b) Assign all the lines assigned to the component candidate to the new repeat component.

\(^1\text{Note that this removes the internal transitivity, which is no longer needed.}\)
The correlation between segments is defined as the correlation between that segment and any other segment. If there are two repeat components that have the same begin and end frame, leave only the one with more lines assigned.

For readability purposes, we do not visualize the repeat segment's occurrences in time and there is no repetition of a repeated segment is itself repeated in a succession, its occurrence is separated by black lines.

For the experiment, the SPPro [7] library was used to implement MFCC computation and Matlab to implement the rest of the algorithm. MFCC feature vectors containing 56 coefficients were generated with sfbcep using 700-ms window size, 100-ms shift size, the Blackman averaging window, the FFT window size of $2^{20}$ and 56 filterbanks. The output was converted to the mel scale and all other sfbcep parameters were set to their defaults.

Binary similarity images were generated using $T_h$ set to 0.625. In order to set the neighborhood size for the LMF algorithm, we assumed the shortest possible repeat component length to be 1 s. Thus, the distance between any two meaningful lines (we assume the meaningful lines are parallel to each other) could not at any point be less than three pixels. The neighborhood for the LMF was then set to 7 (three pixels from each side of the line plus the pixel on the line itself).

L Meincke-processed images were divided into rectangular subregions and the Hough transform was applied on each resulting smaller image, with $\rho$ resolution set to 0.5 and $44.8 \leq \theta \leq 45.2$ (degrees). Hough peaks lower than 7 were ignored, and when a peak was identified, its neighborhood in the peak table was cleaned up (once again, the size of the neighborhood was set to 7). When converting the Hough peaks into a line set, only lines longer than $5\sqrt{2}$ with no gaps longer than $2\sqrt{2}$ were picked up. The lines were adjusted with $\beta_1$ and $\beta_2$ of 4 and 3, as in the reference data the majority of the lines have the angle $41.99^\circ < \theta < 47.73^\circ$. $T_h$ was set to 0.95.

To obtain volume transition information, the RMS of the audio data was computed with a 100-ms window and the output was averaged using Savitzky-Golay coefficients with the neighborhood $k = 100$.

5. EXPERIMENT AND RESULTS

We selected five compositions for the evaluation of our method. We refer to these compositions using identifiers rather than names for the sake of compactness. The list of music titles used along with their identifiers is given in Table 1. We aimed to select music which should be familiar to most readers and for which the score is easily available. ME and BU were all used also in [10] so that the direct comparison between the approaches used is possible. Audio was encoded as monaural PCM data with sampling frequency of 44100 and 16-bit quantization.

The following subsections cover the analysis of the repetition identification and the visualization for each composition. Each subsection starts with a description of the ground-truth (GT) for the tune obtained from its score (the GT consist of the numbers of bars of the score that contain repeated patterns and the time location of the corresponding fragments in the audio data). The discussion about the identified repeat patterns follows. Next, possible interpretations of the misual image for the tune are given together with the discussion on their accurateness. Each subsection includes the misual image for the tune and a table summarizing the identification of the GT repetitions.

5.1 CR

This piano composition was extracted from [12]. According to the score, the piece has four different repeating patterns, in Table 2 marked as #1b-d and #2. The pattern #1b includes #1c, which in turn includes #1d. This nested

4. IMPLEMENTATION

(c) Assign the number of the first frame of the component candidate to the new repeat component. 

Remove lines that end at $x_i$ from all line sets assigned to all candidate components. If for any candidate component the core line set becomes empty, remove it from the list.

4. If there are two repeat components that have the same begin and end frame, leave only the one with more lines assigned.

Use the first frame/last frame information to convert line data assigned to every repeat component to proper timestamps.

3.8 Visualization

Apart from the information about repetitions, various types of other data may be helpful when trying to imagine a composition without listening to it beforehand. Volume, tempo, key, mode, composer, creation date, instruments used, presence and type of vocals are only some of the possible tune characteristics. We decided to integrate our repetition extraction algorithm with the misual framework, which visualizes music by combining the structure and volume transition information. Misual images were successfully used to reason about a composition’s characteristics in [10].

The base image used to visualize the structure consists of a set of 3D-like rings drawn in succession. The position of the center of a ring on the horizontal axis represents the ring’s position in time, while its radius represents the current volume. The volume data is obtained by computing the Root Mean Square (RMS) of the signal and averaging it with Savitzky-Golay weighting coefficients. By fixing the nature of volume transition between tunes.

The repetitions are represented by colored rings on the base image. We distinguish two types of repetitions. A simple repetition occurs when an audio fragment is repeated in succession (the first occurrence is immediately followed by the second and then possibly the next one) and does not belong to any nested repetition structure. A simple repetition is always indicated by light gray with a black strip separating different occurrences of the pattern. Thus, different white ring groups on the image indicate different simple repetitions. In other cases, a separate color is used for each repeated structure.

The internal structure of a repeat component, if any, is indicated by gradation of the base color, the longest subpattern of the component having the darkest color. If a subpattern of a repeated segment is itself repeated in a succession, its occurrences are also separated by black lines.

For readability purposes, we do not visualize the repeat segments that were identified but are unlikely to be noticed by listener. We call the rule used to decide whether a component should be visualized the omission rule. The omission rule says, that a segment is unlikely to be noticed if it is short, its occurrences are separated in time and there is no correlation between that segment and any other segment. The correlation between segments $s_1$ and $s_2$ occurs if occurrences of $s_1$ always precede (follow) occurrences of $s_2$ (see the yellow and pink segments in the Fig. 5).
Table 1: Music pieces used in the paper

<table>
<thead>
<tr>
<th>CR</th>
<th>B95s</th>
<th>VR</th>
<th>ME</th>
<th>BU</th>
</tr>
</thead>
<tbody>
<tr>
<td>F. Chopin, Prelude No.15, op.28</td>
<td>L. van Beethoven, 9th Symphony, 4th movement (fragment)</td>
<td>G. Verdi, Dies Irae from Messa da Requiem</td>
<td>W. A. Mozart, Eine Kleine Nachtmusik, 1st movement</td>
<td>L. van Beethoven, 5th Symphony, 1st movement</td>
</tr>
</tbody>
</table>

group represents the main motif of the composition and the melody returns six times to its initial fragment (#1d). The pattern #2 represents the long, slow melody from the middle of the tune, repeated in succession.

The nested structure #1b-d has been identified with additional nesting level. The algorithm indicated a longer segment #1a including #1b and occurring twice in the composition. The 1.2-second difference between #1a and #1b contains a group of three notes. One of these notes changes its pitch by one tone between pattern occurrences, so although they sound similar, #1a was not part of the GT. The whole repeat pattern is marked with green in the visualization (Fig. 5). Gradation shows the nesting of the components\(^2\). The repeated pattern #2 has been divided because the audio segments (122.8:129.6) and (177.2:184.6) were not identified as similar. The divided fragments are shown in the visualization as light blue, red and dark blue (between the second light blue and the second red). The division might be caused by the large amount of pedaling used in that fragment of the music. The two other two components not indicated by GT are the yellow and the violet one. The yellow fragment contains similar musical phrase, but in its second occurrence the sounds produced by the left hand of the pianist are lower. The violet fragment occurrences also sound similar, but their pitch differs slightly. Both yellow and violet components can be treated as proper repetitions.

When looking at the image, our attention is immediately drawn to its middle part, where a red component is repeated twice each time with the volume slowly rising to become sufficiently higher than the average of the tune. This fragment may be imagined as an especially powerful, central part of the composition. The accompanying volume is always low if compared to the rest of the tune, which may suggest a calm mood building up when the motif is hears. The mutual position relationship between yellow and violet patterns\(^5\) may indicate that they embrace two fragments of the tune musically related to each other.

Listening to the music confirms the powerful character of the central part of the composition and the existence of calm recurring motif indicated by green. The musical phrases located between yellow and violet rings are related to each other, as it could have been deduced from the image.

5.2 B95s

The audio data for B95s was extracted from \[2\]. The excerpt starts two bars before the well-known “Ode to Joy” part of Beethoven’s 9th Symphony finale\(^4\) and contains 54 bars. The only repeat component is the chorus repeated twice as indicated in Table 3 (bars and seconds are counted from the beginning of the excerpt). The repeat component has been correctly identified by the algorithm (Table 3)\(^5\). Misual further indicates, that the composition contains another characteristic element: a frequently recurring motif marked by green. The image shows that the composer (player) returns its first, brightest part six times during the composition. The accompanying volume is always low if compared to the rest of the tune, which may suggest a calm mood building up when the motif is hears. The mutual position relationship between yellow and violet patterns\(^5\) may indicate that they embrace two fragments of the tune musically related to each other.

Listening to the music confirms the powerful character of the central part of the composition and the existence of calm recurring motif indicated by green. The musical phrases located between yellow and violet rings are related to each other, as it could have been deduced from the image.

\(^5\)The same might be said about light blue and red, but as Table 5 indicates, that pair should be merged

\(^4\)Starting with words “Freude, schöner Götterfunken”.

\(^5\)Additionally, a short fragment of the chorus has been connected with a similar fragment at the beginning of the excerpt, not visualized due to the omission rule.

![Figure 5: Misual of CR.](image-url)
Table 3: GT detection results for B95s

<table>
<thead>
<tr>
<th>ID</th>
<th>Bars (＃)</th>
<th>Time (sec)</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19-34</td>
<td>17.3-32.0</td>
<td>OK</td>
</tr>
<tr>
<td></td>
<td>35-50</td>
<td>32.0-47.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: GT detection results for VR

<table>
<thead>
<tr>
<th>ID</th>
<th>Bars (＃)</th>
<th>Time (sec)</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5-2.0</td>
<td>0.3-1.9</td>
<td>OK</td>
</tr>
<tr>
<td>2a</td>
<td>3-10</td>
<td>3.6-15.3</td>
<td>OK white</td>
</tr>
<tr>
<td>2b</td>
<td>Not in GT</td>
<td>Not in GT</td>
<td>OK red</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>15.3-16.8</td>
<td>Not found</td>
</tr>
<tr>
<td>4a</td>
<td>21-24</td>
<td>30.1-36.1</td>
<td>-</td>
</tr>
<tr>
<td>4b</td>
<td>25-28</td>
<td>36.1-41.9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>30.1-31.6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>31.6-33.1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>36.1-37.9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>37.6-39.1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5: GT detection results for ME

<table>
<thead>
<tr>
<th>ID</th>
<th>Bars (＃)</th>
<th>Time (sec)</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>1-55</td>
<td>1.6-90.8</td>
<td>OK red</td>
</tr>
<tr>
<td></td>
<td>90.8-180.0</td>
<td>91.0-180.5</td>
<td></td>
</tr>
<tr>
<td>1b</td>
<td>1-23</td>
<td>1.6-39.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19.8-39.3</td>
<td>18.8-38.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>23.8-45.4</td>
<td>23.6-45.9</td>
<td></td>
</tr>
<tr>
<td>1c</td>
<td>35-42</td>
<td>55.8-70.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>47.0-56.9</td>
<td>46.7-55.5</td>
<td></td>
</tr>
<tr>
<td>43-50</td>
<td>70.0-82.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>69.0-82.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>160.0-172.0</td>
<td>159.5-172.2</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6: Misual of B95s.

Figure 7: Misual of VR. Note: the $y$ axis has been scaled so that details of the right side of the chart are visible.

5.3 VR

The audio data for VR was extracted from [11]. The audio includes orchestral music and accompanying vocals. Five repeating components are suggested by the score (the fifth one is called #2b in Table 4 to highlight its relationship with #2a). The components #1 and #3 consist of repeated orchestra hits (most of the instruments playing a single attack together). The rest of the repeating patterns include orchestra and vocals sounding together.

The structure identification algorithm correctly localized the first orchestra hits succession #1 and the repetition of the initial choral fragment #2. The fragments containing high-pitched vocals from bars 40-42 has been marked as similar with their counterparts from #2a and they are together shown as #2b. The second orchestra hits succession, #3, was not identified possibly because it contains fragments where a non-pitched tympani sounds on its own. The component #4 was not detected fully. The algorithm identified two additional repeated patterns in the end of the composition - one occurring multiple times (white on Fig. 7) and one occurring twice (green). The occurrences of both patterns differ slightly in lyrics and instruments used so they are not listed in Table 4, but the music sounds similar and thus the detection is correct.

The Misual of the composition suggests that the piece starts abruptly and with its full power and stays this way for more than half of its duration. Moreover, the initial fragments of the composition are repeated next to each other, which may increase the initial strong feeling of the music even more. The sudden volume drop (preceded by its short increase) might indicate a dramatic change in the character of the tune at that moment. The visualization shows that the composition ends with patterns repeated multiple times with a delicate crescendo, which may invoke a feeling of anxiety in the listeners.

The confrontation with the music shows, that the nature of the composition has been properly indicated by the visualization, especially when it comes to the powerful feeling given with the beginning of the tune. Also, the significant change of the volume is shown on the image (although on the contrary to what the image might have suggested, the character of the composition itself does not change greatly at that point). The influence of the multiple repetitions that conclude the piece is similar to what could be deduced from the visualization.

5.4 ME

The audio data for ME was extracted from [3]. It contains orchestral music constructed in the sonata form [10]. Table 5 shows the repeat components that should be identified according to the score. The components #1b and #1c are both included in #1a, and one occurrence of #1b is played on its own in the latter part of the tune. The pattern #1a represents the sonata’s exposition.

The algorithm identified all repetitions indicated by the score. It also correctly revealed that both #1b and #1c include shorter repetitions, indicated by different shades of red in Fig. 8. The components marked with the brightest red, white and green are correctly identified repetitions.
related to one another (they contain almost similar melody, but different keys are represented by different colors). The second occurrence of the violet component is hidden under the white rings, which is caused by the melody changing keys in the middle of a repeated pattern. The yellow and blue repeat components in the end of the score sound similar, so their visualization is justified.

The image clearly shows the existence of a long pattern repeated in succession at the beginning of the tune (its repetitions are separated by a wide black ring). During its second occurrence, the beginning half of the pattern is played louder than during the first one, while the ending half is played similarly. The first half is replayed later in the tune, also loudly. This might lead to believe that the first half, occurring three times in the composition, contains the major and possibly most recognizable motif. The second half of the longest repeating pattern may be also easily recognizable due to its repetitive internal structure (with two-level nesting). The repetition marked with light blue at the end of the tune, together with a significant increase in the volume may suggest a short and powerful finishing accent, common for orchestral performances. As noted in [10], a rich repetitive structure might cause the listeners to quickly become familiar with the tune, which in turn may suggest that the tune would be among the more recognizable compositions of its author.

The misual image correctly identified the repeated exposition of the sonata and its part being replayed in the latter part of the tune. They are the most recognizable part of the composition. The beginning of the exposition is indeed played louder in its second occurrence, but the difference is slight and may not be noticeable due to the occurrences’ time separation. The powerful accent in the end of the composition is also indicated by the visualization. Additionally, the misual image revealed significant number of repetitive structures, characteristic to Mozart.

5.5 BU

This orchestral composition was extracted from [14]. Similarly to ME, the music has the structure of a sonata [10]. The component #1a (Table 6) contains the repeated exposition. #1b is contained in #1a and is repeated on its own in the latter part of the composition. The algorithm correctly identified the repetitions indicated by the GT, additionally showing the internal structure of the exposition (shades of red in Fig. 9). Also, multiple simple repetitions occurring throughout the tune have been identified and marked with white. Components visualized as violet, green, light blue and yellow rings are also proper repetitions (not marked with white due to their time separation or internal structure not visible because of the short duration of the components).

Misual indicates that in this piece there is only one long repetition, localized at the beginning of the composition. The similar volume change pattern in both occurrences of the repetition suggests that it is repeated with the same articulation and character, also when its fragment is replayed on its own in the latter part of the tune. The visualization suggests that the composition is build mostly from different short patterns repeated in succession (the tendency is kept also within the long repeat component at the beginning). This may cause the listener to feel that the tune is somehow ordered and predictable. The final part of the composition may leave the listener in an uplifted mood due to constantly increased volume level [10]. This might be further backed up by the yellow component which may suggest a repeated finishing accent in the melody.

Confrontation with the music confirms that misual properly showed the sonata character of the composition by identifying the repeated exposition. It also properly indicated the high number of simple repeat components throughout the tune. The short average length of repetitions may and lead to the situation where even people familiar with the composition, if asked, may not recall their existence. The yellow repeated component is not as close to the end of the composition as the image suggests (due to the scale), so it does not contain the finishing accent but a repetition occurring slightly before it.

This particular misual image shows one of the advantages of using a single color for all simple repetitions. Here, if we wanted to use a different color for every repeated pattern, 11 additional colors would be needed. Due to the limitations of human perception, some colors used for simple repetitions might then look very similar to colors used inside the first repeated pattern (the exposition), making the misual harder to interpret correctly.

6. DISCUSSION AND FUTURE WORK

Aspects of the extraction and visualization results we wish to highlight are covered below.

Algorithm: The LMF algorithm together with the proposed repeat components extraction scheme is characterized by high precision. Two audio segments which do not sound similar were almost never flagged as a single repetition, regardless of the duration of the component (the only occurrence of this situation never made it to the visualization results due to the omission rule). Also, the algorithm is able to identify even relatively short (one second) repetitive patterns.

Advantages of the color scheme used: By using color gradation to indicate patterns that are contained in one another (Figs. 7, 8) misual allows us to analyze music on two levels simultaneously (structure of a single repeated pattern and structure of the whole composition). Listeners can then easily see more complicated structures where subpatterns of repeat components are used also on their own in different parts of a tune. By using the same color for all simple repetitions misual automatically highlights the differences between the characteristic of repetitiveness. For example, the comparison between Fig. 8 and Fig. 9 immediately indicates that

Table 6: GT detection results for BU

<table>
<thead>
<tr>
<th>ID</th>
<th>Bars (#)</th>
<th>Time (sec)</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>GT detection</td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>1-124</td>
<td>1.6-79.9</td>
<td>1.1-79.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>79.9-158.6</td>
<td>79.6-158.1</td>
</tr>
<tr>
<td>1b</td>
<td>37-58</td>
<td>29.4-41.4</td>
<td>28.9-40.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>107.9-119.9</td>
<td>107.4-119.0</td>
</tr>
<tr>
<td></td>
<td>281-302</td>
<td>273.9-286.0</td>
<td>273.4-284.8</td>
</tr>
</tbody>
</table>

Figure 8: Misual of ME
ME is built from repetitive structures with different levels of nesting, while BU contains mostly short musical phrases repeated in succession.

**Tune comparison aspects:** We can use misuals to directly compare the repetition characteristic among tunes to find similarities that are not easily recognizable while just listening. Two observations can be pointed out here. Firstly, misuals indicated that although ME and BU (Figs. 8 and 9) both share the sonata form (long repeated pattern at the beginning of the tune with its fragment being replayed on its own), they differ in the repetitiveness nature. The misual for ME shows a lot of nesting that would ease the remembering process while repetitions in BU, mostly occurring only once, may not have the same property. Secondly, misuals showed that except B95s (which is an excerpt), all the compositions we use, regardless of their form, begin with a fragment that is further repeated. This might be one of the causes of their popularity. We are likely to be more familiar with fragments that are repeated throughout a tune. And if a tune starts with a familiar pattern, we can recognize the composition even while listening just to the first few bars (as we often do, e.g., when looking for a certain piece of music).

**New level of musical structure analysis:** Misual can visualize the existence of very short repetitions played just twice close together (Fig. 9) or short fragments of a longer repetition played on their own in a different part of the tune (Fig. 7). These repetitions and the relations between them may be hard to spot while just listening (because of the time separation of the occurrences) and the similarity may not be evident from the musical score (as they might differ slightly). By revealing them, misuals add a new point of view to the musical structure analysis. The visualizations are also helpful for listeners. By paying attention to fragments of music indicated as repetitions on our images they may discover yet another dimension of familiar compositions.

**Possible improvements:** The study on the possibility of dynamic adjustment of the MFCC parameters might further increase the algorithm resolution and help eliminate the detection problems pointed out in the paper. Also, a module identifying the same melody played in different keys would improve the repetitiveness indication.

### 7. CONCLUSION

In the paper, we proposed a new method of identification and visualization of repeated patterns in music. The method is based on similarity matrices containing MFCC frame-to-frame distances. On these matrices, similar audio fragments appeared as recognizable lines. They were automatically collected and converted into repeated pattern data with the LMF algorithm we proposed and the Hough transform. Then, the results were visualized in a framework called misual, which produces images of music basing on the length of the tune, repeated patterns structure and volume transition data. We proved that misual properly identifies repeating structures and is a useful tool in music structure analysis. With misuals of well-known compositions we also showed that the algorithm reveals repetitions that might be difficult to notice when just listening to music or reading its score.

### Acknowledgments

This research is granted by the Japan Society for the Promotion of Science (JSPS) through the "Funding Program for Next Generation World-Leading Researchers (NEXT Program)", initiated by the Council for Science and Technology Policy (CSTP).

### 8. REFERENCES


