Proceedings of the 3rd International Conference on Music & Emotion (ICME3), Jyväskylä, Finland, 11th - 15th June 2013. Geoff Luck & Olivier Brabant (Eds.)

A SIMPLE, HIGH-YIELD METHOD FOR ASSESSING STRUCTURAL NOVELTY

Olivier Lartillot, Donato Cereghetti, Kim Eliard, Didier Grandjean

Swiss Center for Affective Sciences, University of Geneva, Switzerland
olartillot@gmail.com

Abstract

The structural dimension of music plays an important role in its affective appreciation. One particular aspect is related to the temporal succession of moments, each characterized by particular musical properties. One classical approach in computational modelling of this aspect is based on similarity matrix representations, where successive states are visualized by successive squares along the main diagonal, bearing some resemblance to checkerboards. One referential method estimates a so-called novelty curve, representing the probability along time of the presence of transitions between successive states, as well as their relative importance. Novelty is traditionally computed by comparing – through cross-correlation – local configurations along the diagonal with an ideal checkerboard kernel. The method is limited by a strong dependency on kernel size, which imposes a single level of granularity in the analysis and fails to grasp common musical structures made of a succession of states of various sizes. We introduce a simpler but more powerful and general method that automatically detects homogeneous segments of any size. Only half of the similarity matrix is retained, in order to compare each new instant solely with the past and exclude the future. For each instant in the piece, novelty is assessed by first determining the temporal scale of the preceding homogeneous part as well as the degree of contrast between that previous part and what just comes next. Detailed results show how and why this method offers a richer and more intuitive structural representation encompassing all granularity levels.

Keywords: structure, similarity matrix, novelty

1. Introduction

The structural dimension of music plays an important role in its affective appreciation. One particular aspect is related to the temporal succession of moments, each characterized by particular musical properties. The idea is to automatically segment audio files into a series of homogeneous sections, through the estimation of temporal discontinuities along diverse alternative features such as timbre in particular (Foote & Cooper, 2003).

One classical approach in computational modelling of this aspect is based on similarity matrix representations, where successive states are visualized by successive squares along the main diagonal, bearing some resemblance to checkerboards. The referential method estimates a so-called novelty curve, representing the probability along time of the presence of transitions between successive states, as well as their relative importance (Foote & Cooper, 2003).

We show the limitation of the state of the art and introduce a new method that offers a richer and more intuitive structural representation encompassing all granularity levels.

The model has been implemented in *MIR-toolbox* (Lartillot & Toiviainen, 2007) and is available in the new version 1.5 of the toolbox.

2. Similarity matrix

One common MIR method to describe the structural content of music is based on *dissimilarity* and *similarity matrices*, constructed from a selected audio or musical feature. A *dissimilarity* (resp., *similarity*) *matrix* shows the dissimilarity (resp., similarity) between all possible pairs of frames from the input data. The matrix is constructed as follows: At each successive instant of time (each column *x* in the matrix), numerical distance (in the case of dissimilarity matrix) or numerical similarity (in the case of similarity matrix) is computed between that current instant (point (x,x) on the diagonal of the matrix) and previous instants (points (x,y<x) that are below point (x,x)), as well as succeeding instants (points (x,y>x) that are above point (x,x)).

A graphical representation of the dissimilarity (resp., similarity) matrix, as in Fig. 1, shows these numerical distances (resp., similarities) using a color convention. In *MIRtoolbox*, high values are indicated by warm colors (red, yellow) whereas low values are indicated with cold colors (dark blue, light blue). In a dissimilarity matrix, the main diagonal, representing the absence of dissimilarity between one time frame and itself, is by property dark blue (Fig. 1). Similarly, the main diagonal of a similarity matrix is by property red (Fig. 2).

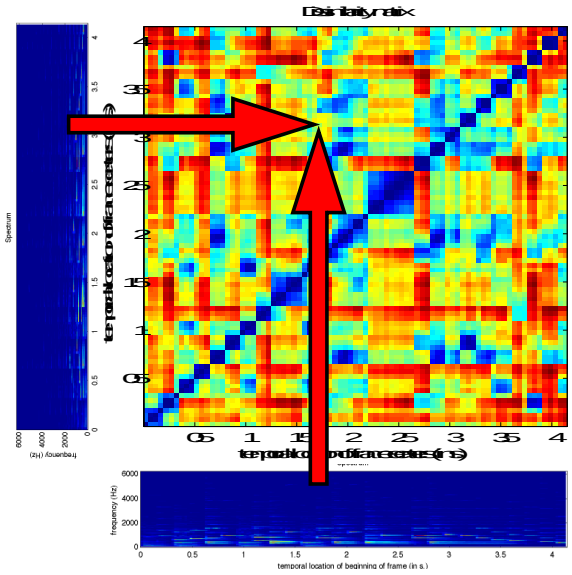

Figure 1. Dissimilarity matrix (top right) showing dissimilarities between frames of a spectrogram (represented both left and bottom).

In the dissimilarity matrix, the distance at each point can be computed using various distance measures. Cosine distance seems to be a good choice for most features, because it enables to compare profile of feature vectors without focusing on the absolute amplitude of the elements of those vectors. For instance, when comparing key strengths (Krumhansl, 1990; Gómez, 2006) between two frames, what is of relevance is the relative importance of tonal centers, not the actual energy in each of those. This distance is chosen my default in *MIRtoolbox*:

ks = mirkeystrength(filename, 'Frame')
dm = mirsimatrix(ks, 'Dissimilarity');

On the contrary, for other features such as Mel-Frequency Cepstral Coefficients (MFCCs), the absolute values are of importance for the comparison. In such case, Euclidean distance, for instance, is more suitable:

cc = mirmfcc(filename, 'Frame')
dm = mirsimatrix(cc, 'Dissimilarity',
'Distance', 'Euclidean');

In similarity matrices, similarity at each point is computed by turning the dissimilarity measure into a similarity measure based on a transformation. One simple choice is a linear transformation of the type *y = 1-x*; a common alternative is an exponential transformation of the type *exp(-x)*, which emphasizes a small dissimilarity through a more important drop of similarity than in the linear transformation. In the following, we will use this exponential similarity measure, which is chosen by default in *MIRtoolbox* when computing the similarity matrix:

sm = mirsimatrix(dm, 'Similarity')

Dissimilarity and similarity matrices reveal homogeneous parts: successive homogeneous states are visualized by successive squares along the main diagonal, bearing some resemblance to checkerboards.

3. Previous "kernel approach" for novelty curve estimation

The main idea behind the notion of *novelty* is that the structure that can be seen from the similarity matrix, with the succession of homogeneous states, could be explicitly represented in a temporal curves where peaks indi-

cate the position of those transitions, and the height of the peaks would correspond to some kind of structural importance.

In a seminal approach (Foote & Cooper, 2003), the estimation of novelty curve is based on the observation that important structural transitions can be seen in the matrix as a succession of squares along the diagonal. For that reason, the novelty curve at each time frame *t* is estimated by comparing the subpart of the matrix around the corresponding point on the diagonal, i.e., (t,t), with an ideal representation of a structural transition, modelled as a 2*2 checkerboard made of two red squares on the diagonal (representing high similarity between a same segment) and two blue squares outside the diagonal (representing low similarity between successive segments). More precisely, the novelty curve results from the cross-correlation of a Gaussian smoothed checkerboard kernel along the diagonal of the similarity matrix.

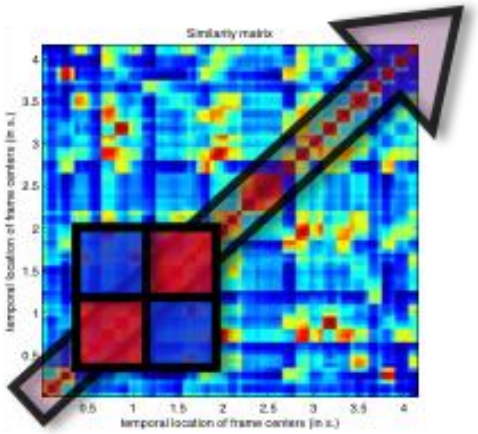

Figure 2. In the kernel approach, segmentations are detected by comparing each successive subpart of the similarity matrix along the diagonal with a checkerboard kernel, idealizing a perfect succession of two segments of same size.

The method is strongly dependent on the specification of the size of the checkerboard kernel. Fig. 4c, 5c, 6c, 7c, 8c and 9c show examples of novelty curves extracted using a kernel of size 64 samples, based on the following *MIRtoolbox* command:

mirnovelty(sm,'KernelSize',64)

Fig. 4d, 5d, 6d, 7d, 8d and 9d show examples of novelty curves extracted using a kernel of size 16 samples:

mirnovelty(sm,'KernelSize',16)

As we will see more in detail in Section 5, this kernel-based method imposes a single level of granularity in the analysis, thus failing to grasp common musical structures made of a succession of states of various sizes.

4. New approach

We introduce a simpler but in the same more powerful and more general method. The idea is to automatically detect homogeneous segments of any size (or *temporal scale*). Future events are excluded in order to focus on the temporal causality of music perception. This means that only half of the similarity matrix, below the main diagonal, is retained.

For each successive column in the similarity matrix, corresponding to a time frame *t*, the novelty value at that time is estimated by detecting whether a homogeneous segment ends just before *t* (cf. Fig. 3). More precisely, the idea is to estimate the *temporal scale* of the previous ending segment as well as the *contrastive change* before and after the ending of the segment. The novelty value is then represented as a combination of the temporal scale and the amount of contrast.

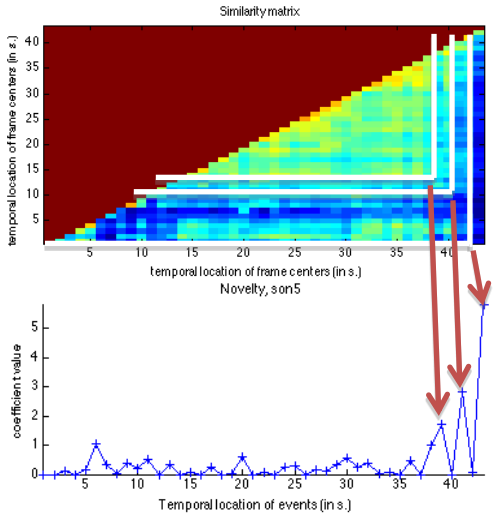

Figure 3. Dissimilarity matrix (top) and its corresponding novelty curve (bottom), computed using the new approach. For the three peaks at time t = 39, 41 and 43 s, the corresponding triangular homogeneous segments are shown in the matrix.

For the particular column at time *t*, in order to assess the temporal scale of the segment ending just before *t*, we consider the triangular

part of the similarity matrix that is below the main diagonal and left to the column *t* (cf. examples of triangular part highlighted in Fig.3). The idea is to detect how much of this triangle, starting from its apex at point *(t,t)*, can be considered as a whole homogeneous segment that is globally of higher value than the new column on its right, at time *t*. The triangle is progressively constructed from its apex *(t,t)* downward, line by line, by checking whether each new line to be added to the triangle is globally of higher similarity values than the next point at time *t*. More precisely, for a given line, we check that both following conditions are fulfilled:

- the new similarity value at time *t* is lower than the similarity value at time *t* – 1 (i.e., the rightmost point of the triangle at that line).
- the new similarity value at time *t* is lower than two standard deviations below the mean of the similarity values along the current line of the triangle.

Once this does not hold true anymore, the construction is interrupted, and we keep the triangle above this unsuccessful line. We obtain a triangle that corresponds to the "ending segment", and the height of the triangle corresponds to the *temporal scale* of this ending segment.

The amount of contrast between this triangular segment and the new column at time *t* is simply computed as the city-block distance between the last column of the triangle and the new column at time *t* for that particular temporal scale.

As we will see in the examples in the next section, this method offers a structural representation that encompasses all granularity levels. This approach is integrated into MIRtoolbox 1.5 and is used by default when calling *mirnovelty*.

5. Detailed analysis of one piece of music

This section presents various structural analyses – spectral (Fig. 4 and 5), timbral (Fig. 6), tonal (Fig. 7 and 8) and metrical (Fig. 9) – of the first 160 seconds of a performance of the *Scherzo* of L. van Beethoven's *Symphony No.9 in D minor, op.125*. Each figure shows first the similarity matrix (Fig. 4a, 5a, etc.) – or more precisely the half, below the main diagonal, corresponding to the memory of the music already heard, with respect to each current time – followed by the novelty curve estimated using the new method introduced in the previous section (Fig. 4b, 5b, etc.), as well as two versions of the kernel-based method using two different granularity levels (Fig. 4c and 4d, 5c and 5d, etc.)

Whereas in the kernel-based method, close points are highly correlated (Fig. 4c and 4d for instance), the curve produced by the new method (Fig. 4b for instance) precisely indicates the temporal location of various segmentations with relatively isolated pulses.

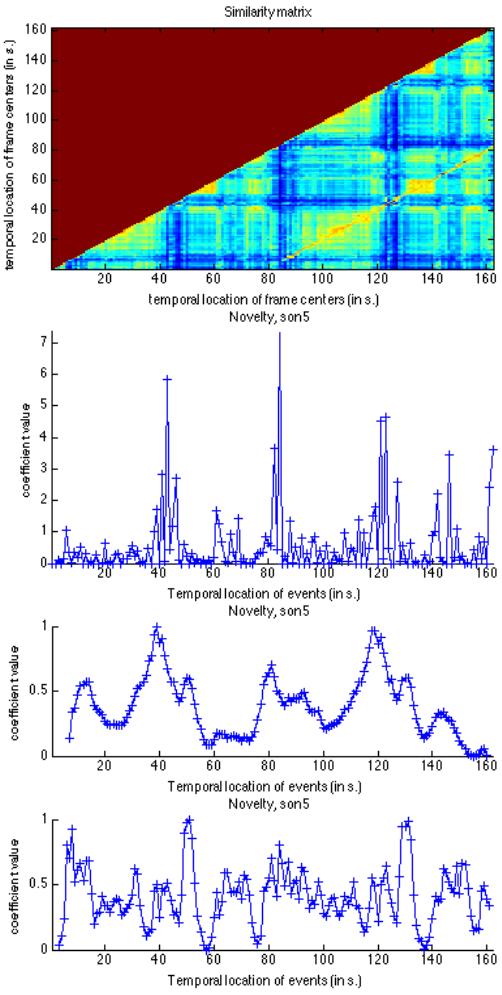

Figure 4. Similarity matrix (4a) and novelty curves based on the new approach (4b) as well as the kernel-based approach with kernel size 64 samples (4c) and 16 samples (4d), all based on a spectrogram with frame size 2 seconds, and a hop of 1 second.

In the kernel-based method, the choice of kernel size has a strong impact on the result. For large kernels, highest peaks in the curve might indicate important segmentation points in the piece, but lower peaks and the curves in-between might be more difficult to interpret. When using a shorter kernel, the actual size of the larger homogeneous parts is not taken into account. In fact, peaks may indicate a transition between segments, an ending segment or a starting segment (such as around t = 55 s in Fig. 5d), which makes the result more difficult to interpret.

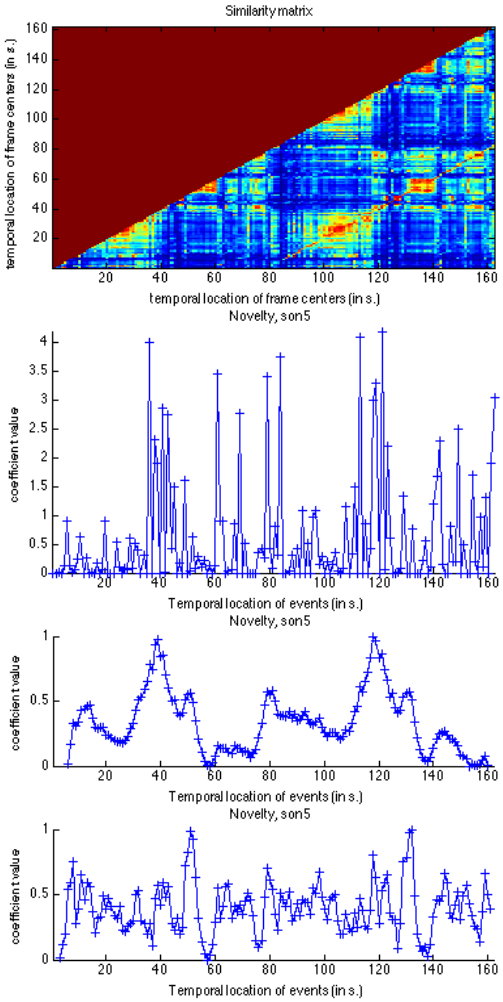

Figure 5. Same as in Fig. 4 but with an autocorrelation function as input, with frame size 2 seconds, and a hop of 1 second.

Fig. 6c and 6d eloquently show a main limitation of the kernel approach: since the idea was to detect structure resembling checkerboard patterns, the transition between two squares of same size, in particular around t = 30 s and t = 110 s, are considered as the most important segmentation points in the piece. The new approach, on the contrary, shows that this transition is not of high importance. It rather highlights the presence of more salient structural endings, related to homogeneous parts of larger temporal scale (such as the 30 s long part ending a little after t = 40 s, t = 70 s, etc.; or the 70 s long part ending at the half and at the end of the piece).

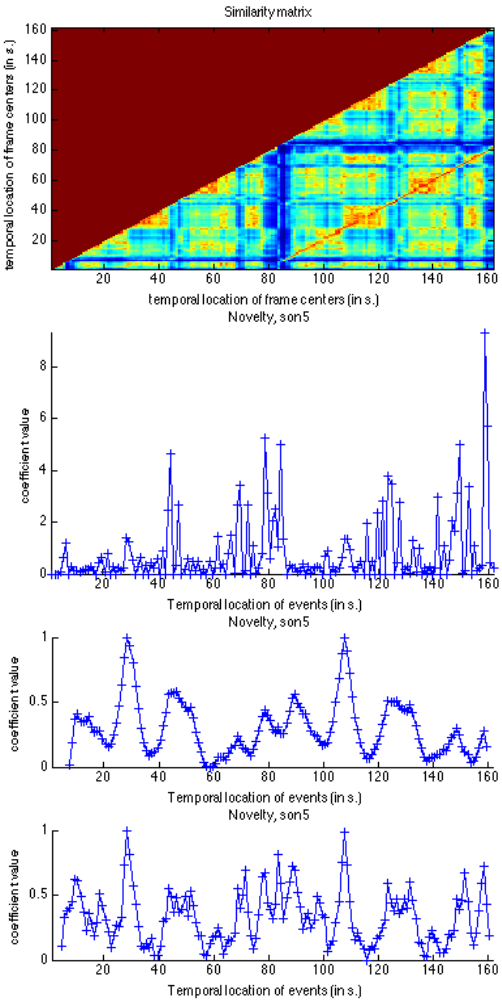

Figure 6. Same as in Fig. 4 and 5 but with MFCCs as input, with frame size 3 seconds, and a hop of 1 second. The dissimilarity is based on Euclidean distance.

On the other hand, Fig. 7b shows a limitation of the current version of new approach, due to a hypersensitiveness to isolated columns that are highly dissimilar to neighbor columns, such as around t = 90 s. This problem can be avoided by filtering out somewhat the isolated column through blurring: by considering a frame size of 5 s instead of 2 s (Fig 8a and 8b), the isolated columns are less important,

and not detected anymore by the new approach. We notice also that these outlying columns do not affect the kernel-based approach (cf. Fig. 7c and 7d, compared to Fig. 8c and 8d).

This shows that the current version of the new approach cannot properly handle similarity matrix with highly salient isolated columns.

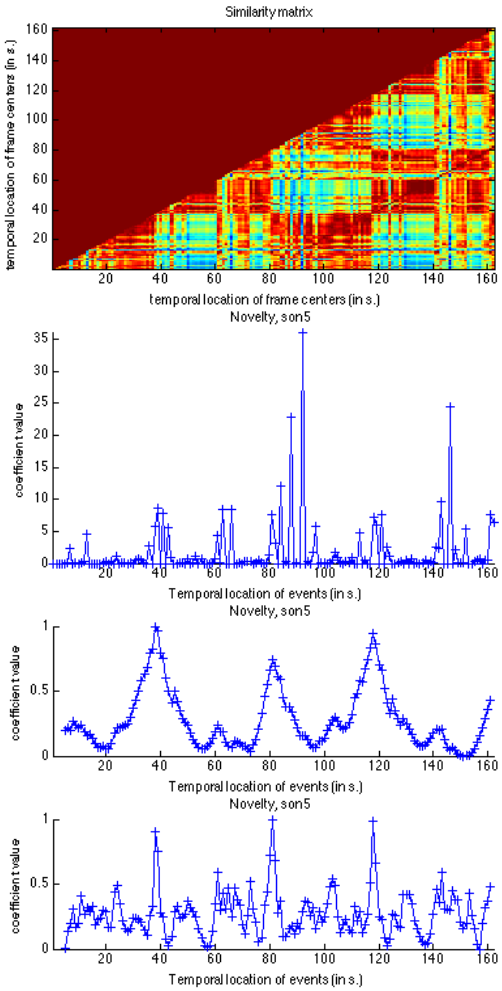

Figure 7. Same as in Fig. 4 to 6 but with key strength vectors as input, computing on a 2-second long moving window with 1-second hop.

The approach might be made more robust by finding more suitable conditions governing the construction of the triangle presented in section 4, in order to better treat such outlying columns and lines in the similarity matrices.

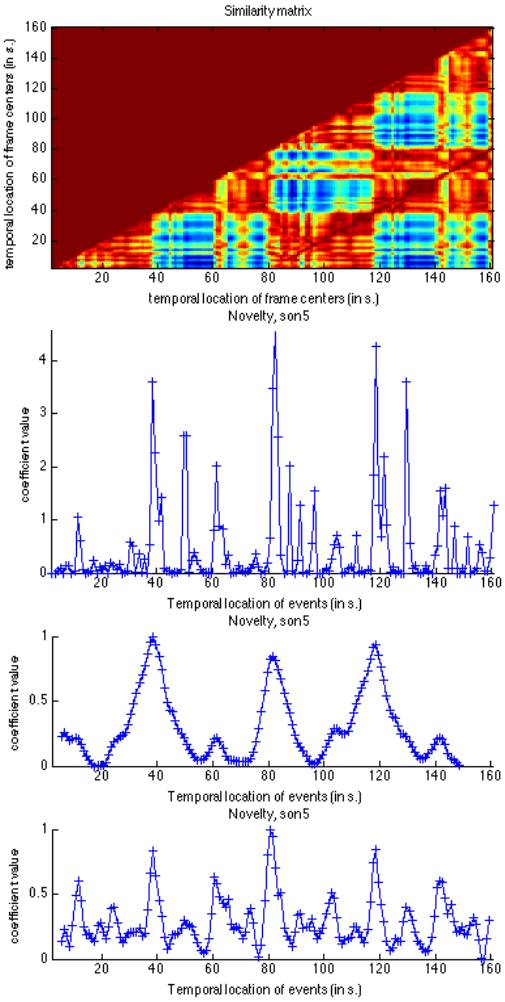

Figure 8. Same as previous figure but with a window length of 5 seconds.

Figure 9 shows another interesting property of the new approach for estimating novelty curve. The transition from one homogeneous state to the next one can sometimes be progressive, and each state can progressively decay over several frames, as can be seen in Fig. 9a with the progressive gradient of colors from dark red to dark blue at the right end of the triangular parts. This happens in particular when the frame decomposition is based on a smaller hop, such as .25 s in Fig. 9. In such case, the new approach for novelty curve shows not a single pulse, but a lobe, still very sharp, but with a width of several frames. Hence in the new approach, the importance of peaks in the novelty curve is indicated not only by height, but more generally by the area of such sharp lobes.

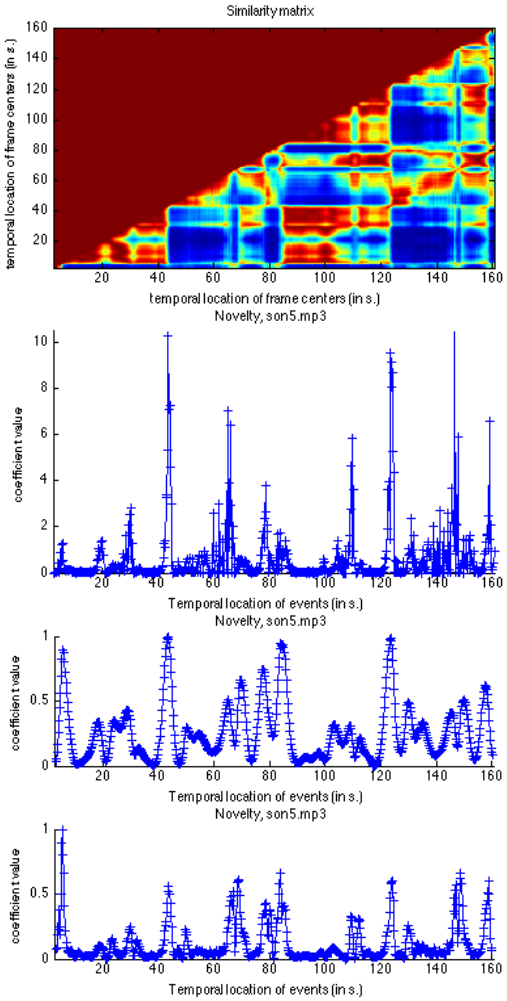

Figure 9. Same as in Fig. 4 to 8 but based on a metrical autocorrelogram with a hop equal to .25 s. (The first half of this metrical autocorrelogram is shown in Fig. 6a in (Lartillot et al., 2013).)

As for a more qualitative and musical conclusion concerning the structural analysis of this performance of Beethoven's *Symphony No.9 Scherzo*, we can notice the very interesting differences and imbrications between the similarity matrices computed from the different audio and musical features.

6. Synchronicity between novelty curves

In this paragraph, we evaluate the synchronicity of the novelty curves computed from main audio and musical features, using the current version of the algorithms while writing the paper, i.e., *MIRtoolbox* 1.4.1.5. The structural analysis is based on the following features:

- spectrogram with frame size 2 seconds, and a hop of 1 second:

a = mirspectrum(filename, 'Frame', 2, 's', 1, 's')

- cepstrogram with frame size 2 seconds, and a hop of 1 second:

b = mircepstrum(a)

- autocorrelation function of the audio waveform with frame size 2 seconds, and a hop of 1 second:

c = mirautocor(filename, 'Frame', 2, 's', 1, 's')

- MFCC with frame size 3 seconds, and a hop of 1 second:

d = mirmfcc(filename, 'Frame', 3, 's', 1, 's')

- chromagram with frame size 2 seconds, and a hop of 1 second:

e = mirchromagram(filename, 'Frame', 2, 's', 1, 's')

- key strength with frame size 2 seconds, and a hop of 1 second:

f = mirkeystrength(e)

- chromagram with frame size 5 seconds, and a hop of 1 second:

g = mirchromagram(filename, 'Frame', 5, 's', 1, 's')

- and key strength with frame size 5 seconds, and a hop of 1 second:

h = mirkeystrength(e)

All features were extracted from thirty-six musical excerpts covering a large range of musical styles from baroque to contemporary classical music (Eliard et al., 2013; Eliard & Grandjean, in preparation) with a mean duration of 155.83 ± 10.66 seconds. As indicated in section 2, similarity is based on cosine distance for all features:

n1 = mirnovelty(a), *n2 = mirnovelty(b)*, etc.

except for MFCCs where Euclidean distance is used instead:

n4 = mirnovelty(d, 'Distance', 'Euclidean')

In order to evaluate the similarities between the different novelty curves, we compute a normalized cross-correlation – without centering – between each pair of novelty

Table 1. Normalized cross-correlation – without centering – between each pair of novelty curves computed from the following features: spectrogram (spec), cepstrogram (ceps), autocorrelation function (ac), MFCCs (mfcc), chromagram with 2 second frame (chro2) and key strength (keys2), and same for 5 second frame (chro5 and keys5).

	spec	ceps	acor	mfcc	chro2	keys2	chro5	keys5
spec	1	.42	**.78**	.41	**.7**	.45	.25	.19
ceps		1	.35	.35	.35	.26	.25	.18
acor			1	.36	**.65**	.45	.22	.16
mfcc				1	.3	.23	.26	.21
chro2					1	**.63**	.21	.15
keys2						1	.17	.12
chro5							1	**.66**
keys5								1

curves. Due to the particular aspect of the novelty curves given by the new approach, where the presence of isolated peaks makes the distribution non-Gaussian, it would not make sense to assess the linearity between curves based on Pearson correlation. We note however that due to the fact that the novelty values are always positive, and that most values are low and very few are high, a direct cross-correlation between novelty curves will show whether their peaks are well synchronized or not. The cross-correlation can be normalized in the same way as a traditional cross-correlation, except that in our case there is no centering, since the absolute magnitude (whether a point belongs to a peak, or is close to zero) plays an important role.

The results of the correlations are shown in Table 1. We can see that spectrum-based and autocorrelation-based novelty curves are highly similar, and similar to chromagram with same 2 s long frames. This may be intuitive since spectrum and autocorrelation function are two different but closely related low-level description of audio, and that chromagram is directly based on spectrum. Chromagram and key strength with same frame size are highly related, because keystrength is highly based on chromagram. On the other hand chromagrams (or keystrengths) with different frame sizes are not cross-correlated at all. This may be due to the problem related to excessive peaks using the 2 second long frame, as discussed in section 5.

7. Discussion

As explained in section 4, in the new approach, the novelty values correspond to a combination between two factors: the *temporal scale* of the previous ending segment and the amount of *contrastive change* before and after the ending of the segment. We might consider in future works a study of each factor separately, and a study of the optimal combination between these two factors.

We also observed in section 5 that in high-resolution similarity matrices, the new approach for novelty curve might include not solely single pulses, but also sharp lobes with a certain width. We noted that the importance of peaks in the novelty curve is indicated not only by height, but also more generally by the area of such sharp lobes. An alternative representation would be to integrate the novelty curve, so that the obtained novelty values would indicate the total contrastive change before and after a progressive transition between segments.

References

Eliard, K., Cereghetti, D., Lartillot, O. & Grandjean, D. (2013). Acoustical and musical structure as predictors of the emotions expressed by music. *Proceedings of the 3rd International Conference on Music & Emotion*, Jyväskylä, Finland.

Eliard, K. & Grandjean, D. (in prep). Dynamic approach to the study of the emotions expressed by music.

Foote, J.T., & Cooper, M.L. (2003). Media segmentation using self-dimilarity Decomposition. *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, 5021, 167-75.

Gómez, E. (2006). *Tonal description of music audio signal*. Phd thesis, Universitat Pompeu Fabra, Barcelona.

Krumhansl, K. (1990). *Cognitive foundations of musical pitch*. Oxford UP.

Lartillot, O., & Toiviainen, P. (2007). MIR in Matlab (II): A toolbox for musical feature extraction from audio. *Proceedings of the International Conference on Music Information Retrieval*, Wien, Austria.

Lartillot, O., Cereghetti, D., Eliard, K., Trost, W.J., Rappaz, M.-A., & Grandjean, D. (2013). Estimating tempo and metrical features by tracking the whole metrical hierarchy. *Proceedings of the 3rd International Conference on Music & Emotion*, Jyväskylä, Finland.