Retrieving and Recreating Musical Form

Ole Kuhl and Kristoffer Jensen

Aalborg University Esbjerg, Niels Bohr Vej 8,
6700 Esbjerg, Denmark
{ok,krist}@aaue.dk

Abstract. This paper discusses musical form from a cognitive and a computational viewpoint. While several time-windows exist in the brain, we here put emphasis on the superchunks of up to more than 30 seconds lengths. We compare a strategy for auditive analysis based on human cognition with a strategy for automatic analysis based on feature extraction. The feature extraction is based on the musical features rhythm, timbre and chroma. We then consider the possible consequences of this approach for the development of music generating software.

Keywords: Music retrieval, human cognition, chunking, feature extraction, music generation.

1 Introduction

If you look through your music collection, you are bound to find that most of the music in it – perhaps 80–90 % – is structured in such a way that a formal change takes place every 30–40 seconds or so. A formal change can be a change from verse to refrain, from A-section to B-section, a repetition, a change of key, etc. Music is generally made up of sections, and the longer time span of the whole piece of music is subdivided into sections with different qualities. This is true for most of the world's musics, regardless of culture and style.

This way of structuring a piece of music is so ubiquitous that it is reasonable to assume that it reflects a built-in characteristic or constraint of the human mind/brain. Such a constraint may well be biologically determined, something that can be deduced from the fact that nursery rhymes all over the world share the same basic structure and the same temporal dimensions [1]. However, this innate tendency of human cognition to structure and group musical sound into sections of certain proportions is difficult to explain. It may be tied to the limitations of our working memory as suggested by some [2, pp 49–51]; or it could be seen as the product of an attention cycle, that would then be the result of the need of the human brain to perform an attention switch every so often in order to reorganize its content [1].

On the computational level, much interest has been put into the automatic segmentation of music into e.g. chorus/verse. The automatic segmentation can be used for many purposes, including creation of a shorter preview with no repetition of chorus, skipping of intro in live DJ situation, for live recomposition, and as an aid in music analysis. In this work, a method for automatic segmentation of music, based on

R. Kronland-Martinet, S. Ystad, and K. Jensen (Eds.): CMMR 2007, LNCS 4969, pp. 263–275, 2008.

features related to the perception of music is used as the basis for a shortest-path method to find segment boundaries. The performance of the features, *rhythmogram*, *timbregram*, and *chromagram*, are then compared to the musical analysis, and the theories from cognitive science. In all cases, the typical segment length is observed and compared. An informal analysis of the musical changes that create the segment boundaries is performed. The result of this analysis is used in a simple, stochastic-based melody generator, and simple changes to the possible notes, the dynamic level and interval between notes is inserted in order to create music and compare the structural changes of this music to that of the musical examples.

2 Temporal Cognition and Musical Form

The temporal organization and function of human cognition is full of complexity. In spite of recent advances in the technology for brain studies we still know very little about the brain's performance over time. Recently, however, the theory of chunking has gained some momentum [2 pp. 47–59, 3, 4 pp. 103–113]. According to this theory, our temporal cognition is structured in three distinct layers, serving different purposes and engaging different brain areas. At the microlevel we perceive the world pre-consciously as perceptual qualities, sometimes called qualia, which are organised in coherent structures, in order to be interpreted or conceptualized. These chunks of information are presented at the mesolevel, where we consciously consider objects and events, statements, gestures etc. In order to bring coherence into the flow of events, we organize the chunks in larger groups or super-chunks at the macrolevel, placing the individual chunk in a larger context.

Psycho-physical evidence shows that the brain has a number of distinct time-windows that can be seen as biological constraints on the cognitive processes [1]. Thus, the pre-conscious microlevel of subchunks extend from 30 ms to 300 ms; the conscious mesolevel of chunks from 300 ms to 3 sec; and the reflective macrolevel of superchunks from 3 sec to roughly 30–40 sec., where the limitation of our memory systems sets in. Naturally, we are not consciously aware of these temporal dimensions, as the brain has developed mechanisms to deal with them so that we can experience the world as a uniform flow of time.

What interests us here is the grouping of chunks into superchunks, which we see as a way of understanding the formal level in music. Individual melodic phrases or gestures are grouped together in superchunks that are limited by the brain's memory capacity. At a simple, generic level, music is organized in sections, the sizes of which fall inside certain boundaries. In fig. 1 below we see the organization of a typical popular song [5] in A-sections and B-sections, further elaborated with intro, outro and a contrasting C-section. The A-sections have a length of 32–33 seconds, while the B-sections are 30 seconds long.

Before we proceed let us note that the definitions of musical form applied here rules out certain highly developed artistic forms, repetitive forms, etc. that cannot be adequately dealt with inside these simple paradigms (see [4 pp. 20–28] for a discussion).

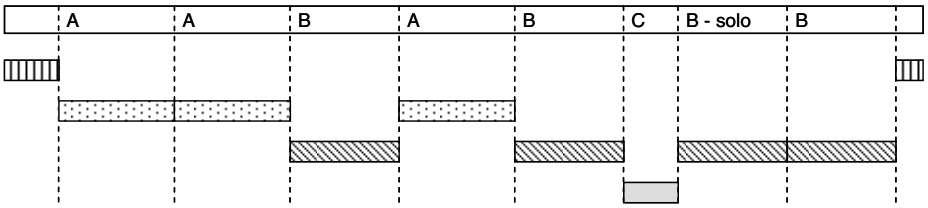

Fig. 1. Jamie Walters: *Hold On* (from [4], used with permission from Peter Lang)

2.1 Musical Form between Change and Continuity

The question we need to answer at this point is: how are the formal properties of a musical piece inscribed in the music-as-sound? Let us look at fig. 1 again. In a case like this, the form of the piece is defined by the song. The recurring refrain (B-section) and the developing verse (A-section) together determine the main aspects of the formal structure. We can imagine certain variations of the form, and indeed musicians make this kind of variations for live-versions, cover-versions etc. But the contents of the verse and the refrain cannot be changed much without jeopardizing the identity of the piece. In a song like this, the form is mainly determined by the text.

But let us set aside the question of the text for now and focus more closely on the music itself. Even without the text, for instance in an instrumental version of the same piece, the form would be clear. This is possible because the music is designed to enhance the individual qualities of the A-section and the B-section respectively, in order to emphasize the contrast between them. A number of musical parameters are shaped to this end. For instance the singer performs at an intimate medium level – more or less with a speaking voice – in the A-sections (verse), while he sings out at a higher pitch in the B-sections (refrain). Another important point is the contrast between the harmonic structure of the two sections. Also the rhythmic effect of the accompaniment and the dynamic level of the two sections differ considerably. Orchestration is yet another favourite parameter for arrangers, in this case the chorus accentuates the dynamic level of the B-section, when it joins in the *Hold On* refrain.

In other words, we have an example of musical form as established through the balance between continuity and change. Continuity is constituted by the text and its narrative; by the singer's voice and the acoustic space provided by his band (acoustic guitar, bass and drums, organ, chorus and lead-guitar is a safe and well-known frame for a narrative); and by the tempo, which makes us entrain to a certain pulse. Variation is set up in this case through the devices listed above: the level of the voice; harmonic structure; dynamic level; and orchestration. In short: some musical parameters are kept constant, while others change from section to section.

Let us look at another example (fig. 2). In the third movement of Mozart's piano sonata in A major, K. 331, also known as *Alla Turca* [6], there is no song, and consequently no text, to determine the formal division of the piece. The structure is constituted on purely musical grounds, yet we find a form comparable to the form of *Hold On*.

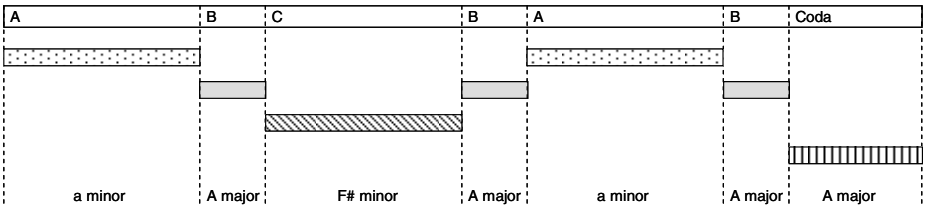

Fig. 2. Mozart's Alla Turca (from [4], used with permission from Peter Lang)

Here the A-sections last 41–42 seconds, the C-section 40 seconds, while the B-sections, that function as refrains and provide the transitions from A to C and vice versa, only last 13–14 seconds. (These timings depend of course on the tempo of the specific performance, here we have used Daniel Barenboim's recording). Continuity is established through a number of parameters: the sound of one piano playing (timbre); tempo and meter (rhythm); key and harmonic development (chroma); repetition and motivic development are some of the more obvious.

As for the parameters used in creating contrast between sections, we can take harmony and texture as evident examples. Harmonically, the A-sections are in A minor, the B-sections in A major and the C-section in F# minor. By texture we mean a combination of rhythmic and gestural qualities, timbre and density. Here the A-section presents a simple melodic gesture with an everyday 'um-da-da-da' rhythm in the left hand accompaniment. The C-section uses the same type of accompaniment, while the right hand abandons the melodic gestures and instead engages in streams of energetic 16^{th}-notes. And the B-section, serving as a transition space between A and C, uses full chords, percussive rhythm and a high dynamic level, thus balancing out its relative brevity with more sound, so to speak.

We shall not go into more details concerning Mozart's musical architecture. The main points to be gained from these examples are, once again, that musical form is constituted through the division of the musical timespan into sections of a certain size; that the individuality of these sections is brought about through a balance between change and continuity; and that this play of variation inside a frame of overall unity is grounded on the tendency of the human mind to create coherence in event structure, which prompts us to generate expectations that can be fulfilled or disappointed.

The purpose of this study of musical form was to investigate the possibilities for a computer-based retrieval of musical form and a modelling of this form, with a subsequent regeneration of music. We will engage the question of whether the musical parameters involved in the human construal of musical form can be defined in such a way that they can be employed in a mathematical analysis of sound. In the following we shall introduce some strategies for retrieval of information pertaining to musical form from the sound stream. We will concentrate on the transition points between sections (A-B; B-C etc), as they mark the points in time, where the change of musical parameters is heard, and where fulfilment/expectation will be experienced. At this stage we have chosen to examine three parameters: timbre, chroma and rhythm.

3 Retrieving Structure in Music

It is our aim to find the same music structure through automatic segmentation as is found through music analysis. This supposedly corresponds to the temporal laws that are the result of temporal processing in human cognition.

We present methods for obtaining measurements of the music that correspond to the sound of the instruments playing (timbre), the tempo and meter (rhythm), and the key and harmonic development (chroma). In order to do this, we first identify the single musical events (notes). After this, the rhythmic feature is found by comparing the single events over time, the timbre feature is found by measuring amplitude over time and frequency, with perceptually related frequency and amplitude features. Finally the chroma is found by summing all partials into the twelve chromas. Both the timbre and chroma features are smoothed over time, to remove small irregularities that occur in addition to more consistent musical events. Finally, we present an approach to the problem of retrieving structure in music.

3.1 Feature Extraction

Notes are the fundamental events in the music considered here. A note has a starting point, a rather short attack, a sustain/decay, and a release. Because the attack generally is short it will be possible to measure the point of attack by calculating the amplitude as a function of time, and taking the time derivative of it. Here, this is done by subtracting the previous time step amplitude from the current time step amplitude. The maximum of the time derivative has been shown to be an important cue when investigating the perception of the attack [7]. By estimating all partial amplitudes, and summing the time derivative of all of them, multiplied with a frequency dependent weight in order to have perceptually normalized amplitudes, a well-performing feature, called the perceptual spectral flux (*psf*), of the note onset detection problem is obtained [8].

The rhythm feature, called *rhythmogram*, is obtained by calculating a windowed autocorrelation function on the *psf*, in which the regularity of the note onsets intervals are found in overlapping segments of the music [8]. The timbre is calculated using a front-end (acoustic pre-processor) used in speech recognition, the perceptual linear predictive (*plp*) analysis [9]. In order to remove noise and intermittent events, the *plp* is smoothed over time [10] using a Gaussian weight. The resulting feature is called the *timbregram*. Finally, the key and harmonic development is measured using the chroma that maps the partials into twelve bands, corresponding to the twelve notes of one octave. The resulting measure, smoothed in the same manner as the timbregram is called the *chromagram*. These three features are visualized in fig. 3. The *rhythmogram* has rhythm interval in seconds on the y-axis, the *timbregram* has frequency in bark [11], corresponding to the perceptual frequency scale, and the *chromagram* has note pitch names. All three have time in seconds on the x-axis.

We can read from the rhythmograms that *Hold On* has a more steady beat, while *Allaturca* varies more in tempo, and even has sections with no clearly distinguishable rhythm. The timbregram also contains the loudness, and illustrates remarkable similar structure in the two songs, with two short crescendos in the first two-thirds of each song, and a longer crescendo in the last fourth. Comparing the two *chromagrams*, the

most striking difference is that of clarity. The Mozart piece presents a clear and simple tonal structure, while the Walters song seems more ‘muddled’. This reflects the fact that Mozart is played on a single well-tuned piano, while the pop song combines several instruments, including drums, all with large profusions of overtones, and furthermore the style itself is defined by lots of micro pitches, melodic glides etc.

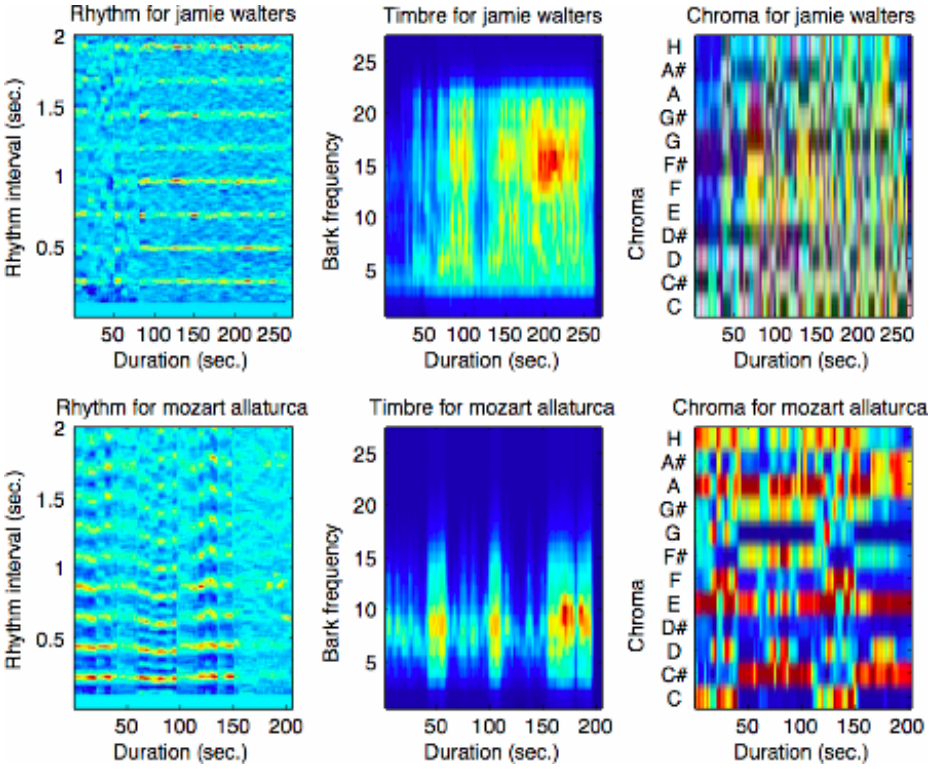

Fig. 3. Rhythmogram (left), timbregram (middle), and chromagram (right), for Jamie Walters *Hold On* and Mozart *Allaturca*. Blue corresponds to little energy and red to much energy.

3.2 Automatic Segmentation

Through a visualization of the extracted music features it is made clear that segmentation of the music should be done in time areas where the feature is homogenous (has the same shape). This is done using the self-similarity measure, originally called recurrency plots [12], which measures the similarity of all the time segments to each other. In fig. 4 the self similarity (calculated as the L_2 norm) is visualized for the same songs and features as in fig. 3.

In the self-similarity plots, the homogenous segments are easily seen climbing the diagonal, in blue/dark. Certainly, the *timbregram* has more homogenous segments than the other two features.

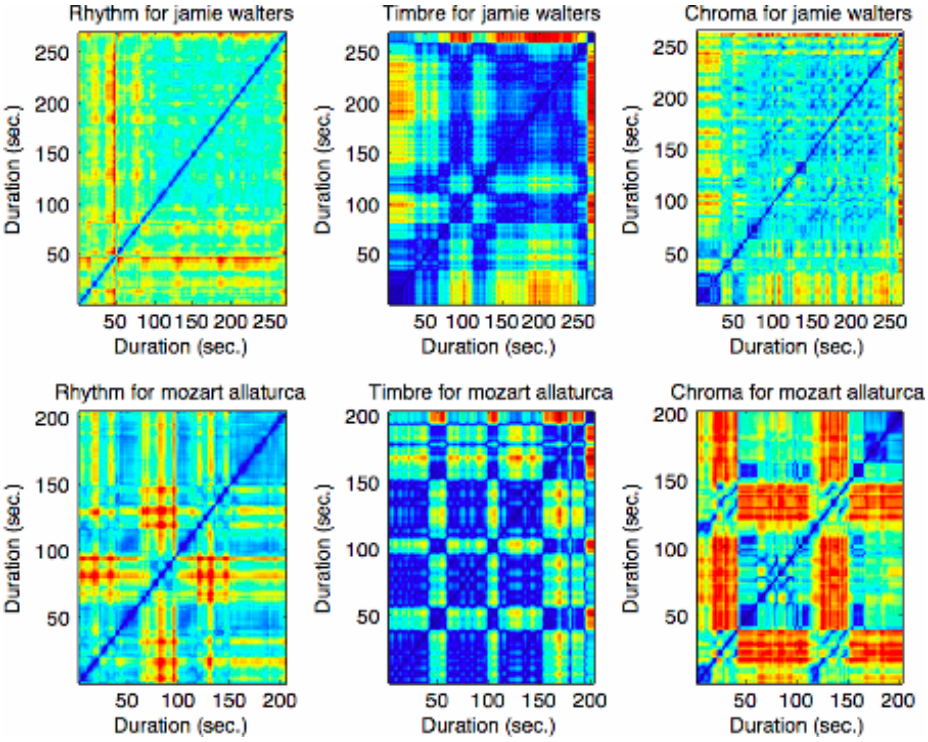

Fig. 4. Selfsimilarity for the rhythm, timbre and chroma of Jamie Walters and Mozart

Several methods exist for identifying the segment boundaries from the self-similarity features. The novelty measure [13], calculated using the checkerboard kernel, gives an indication of the degree of novelty in the audio. If larger segments are to be identified, it necessitates a larger window in the self-similarity matrix. In order to diminish the calculation demands, [8] suggested to smooth the novelty measure calculated on small windows. [14] introduced a shortest path method for finding the optimum path. This method can create segmentation boundaries in many scales, from short segments down to note level, to large segments up to chorus/verse level. This was used in [10] to show that timbre performs slightly better than rhythm or chroma when comparing manual segmentation points with automatic ones.

3.3 Retrieving Structure in Music

In the preceding sections we have introduced two distinct strategies for the analysis of musical form. One is a listening strategy, which is based on the temporal properties of human cognition. The other is a mathematical approach performed automatically by the computer. They both deal with what is almost the same thing: the first with music as sound, the other with sound files. The question we now wish to address is whether these two strategies can yield comparable results.

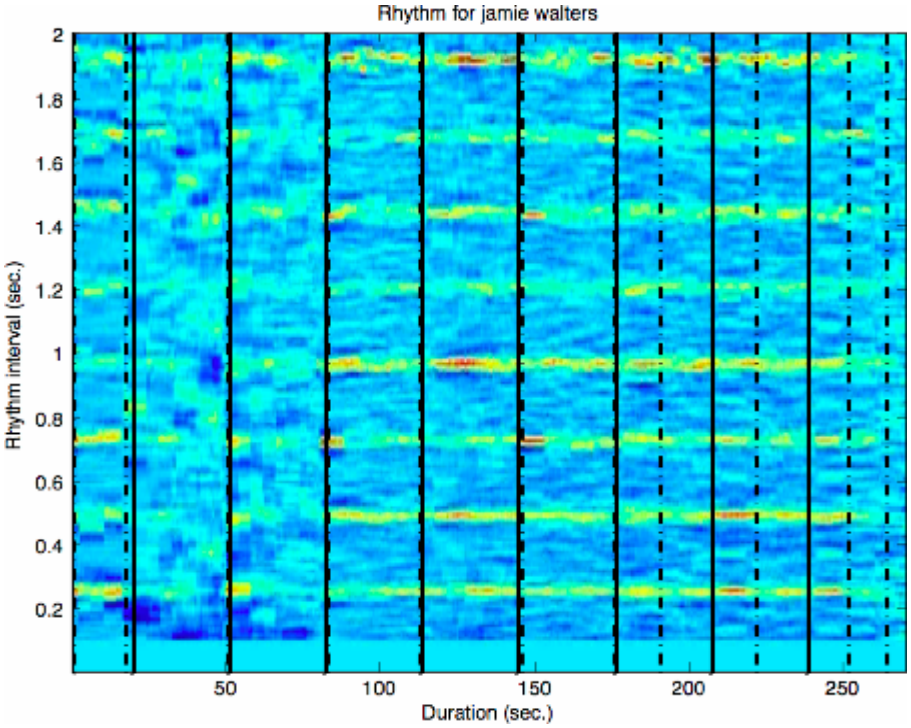

Fig. 5. Rhythmogram for Jamie Walter with automatically found boundaries (solid) and manual boundaries (stipled)

In the human listening strategy we look for changes in texture, orchestration, density, etc., while the computer establishes *rhythmograms*, *timbregrams* and *chromagrams*. Assuming that the extracted musical features, rhythm, timbre and chroma, can be interpreted as essential aspects of such musical parameters as harmony, texture, dynamics, orchestration etc, we would expect to see a correlation between the two approaches. In the following we shall be looking for the transition points between formal sections, such as A-B, B-C etc. These points mark the change from one section with a certain configuration of musical parameters to the next section with a contrasting configuration of musical parameters. They can be easily established by measuring the timing of the two pieces we have studied. The next step will be to insert these points in the *rhythmo-*, *timbre-* and *chromagrams* in order to compare them to the automatic segmentation points. Comparison between the formal and automatic segmentation using the shortest path method is shown in fig. 5 for the Walters *rhythmogram*, and in fig. 6 for the Mozart *chromagram*.

In the Walters *rhythmogram* (fig. 5) we found 11 manual boundaries and 9 automatic ones, with 7 matches between the two. In fact, there is perfect matching up to the point where the C-section, which is only half the length, is introduced. The contrast between B and C is mostly established through timbre and chroma and seems not to be discovered by the *rhythmogram*.

In the Mozart *timbregram* we found 8 manual and 11 automatic boundaries, and 8 of these (all the manual!) match. The three 'extra' boundaries calculated in the *timbregram* might be explained as the result of a barely noticeable playing strategy, in which dynamic contrasts between sections will be enhanced by the player.

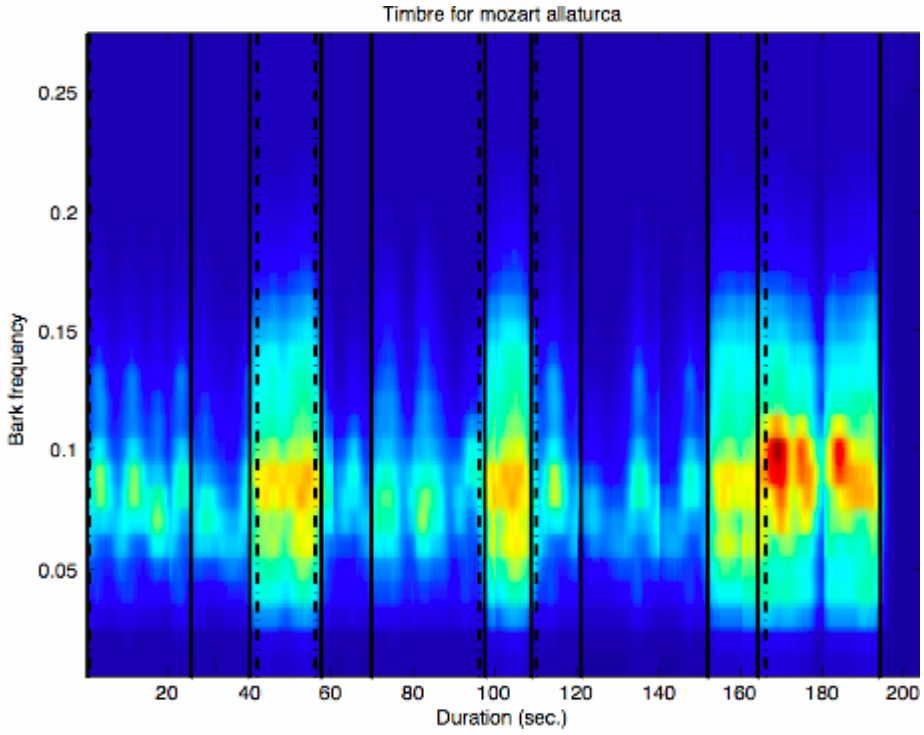

Fig. 6. Timbregram for *Mozart* with automatic (solid) and manual (stipled) segmentation

The *musigrams* visualized in fig. 5 and 6 are the best features for the two songs. The matches correspond to the standard information retrieval measures *recall* (64%, 100%) and *precision* (78%, 73%) and the combined F_1 measure of (0.7, 0.84), the best of the features, with F_1values of (0.7, 0.54, 0.54) for *Hold On* and (0.52, 0.84, 0.7) for *Allaturca* for the *rhythmo*-, the *timbre*-, and the *chromagram*. It is interesting to observe that the rhythmogram performs better with the rhythmic song, while the timbregram performs better with the classical piece.

4 Perspectives in Music Generation

Can the findings in the previous sections be used here to improve music generation algorithms? An attempt to do so is embarked upon here, by using synthesis of melodies made from random notes, using probabilities obtained from music databases. As we wish to take into account the temporal structure of generic music,

we will briefly present some methods to introduce structuring in the rhythm, timbre and chroma of synthetic music, and present some preliminary results.

4.1 Stochastic Models

The stochastic (random) models are implemented in practice by creating a probability density function (*pdf*), which states the probability of an event occurring, as a function of the variable. For instance, one *pdf* could give the probability of one note occurring as a function of the note value (pitch). In tonal music the notes of the musical scale used would typically have a larger probability. The values of the notes can then be found, in music generation, by using the inverse image formula [15].

As an example, the probability of a note being played has been estimated from a database of folk songs[1]. Music generation from simple note probability does not in itself render interesting music, but a conditional probability of the following note, when a note is played, improves the situation. This means that we add the interval probability to the note probability. The probability of an interval is assumed to be independent of the note; therefore the two probabilities can be multiplied in order to create the *pdf* used to find the next note. In this case, a rather pleasant stream of music is created, but still without enough structure to be really interesting (unless perhaps in relaxation use).

4.2 Structural Improvements

The first (*chroma*) improvement to this model is to use a subset of the notes at each structure. Indeed, from fig. 3 it is clear that only 3-5 chromas are played at the time, and from the figure and the discussion in the previous section, this set can be expected to change approximately every 30 seconds. Additional observation of the *chromagram* of 50 songs of different genres [10] reveals a similar chroma evolution for most songs, with variations in number of prominent notes and the rate of change of these notes. Nonetheless, the behavior found in the two songs analyzed here is still the most common. Therefore, this knowledge is inserted in the model, by only choosing a maximum of 5 notes initially, and replacing, adding or removing one note, or doing nothing every N seconds with equal (1/4) probability. N is a random variable with uniform probability between e.g. 30 to 40 seconds, the size of a 'super-chunk'.

The second (*timbre*) improvement is found by looking at the *timbregram* in fig 3. Indeed, both the rock and the classical music display the same structure, with respect to the timbre; a more quiet part is replaced by a louder part, this is repeated, followed by an even louder part and finally, the songs are ended by a strong segment. The strong parts seem to have relatively more energy in a higher frequency range. The quiet part is supposed to correspond to a verse, and the louder part to the chorus. Again, further observations of the *timbregram* of 50 songs [10] broadens the picture, as the number of chorus/verse repetitions, seems to vary between one and eight, the number of verses preceding the chorus vary between one and four, the dynamic difference between the chorus and the verse is also different between songs, and varying other differences are found, including more or less prominent intro/outros.

[1] The Spring 2002 Digital Tradition Folksong Database, http://www.mudcat.org, (1 Nov. 2007).

A simple improvement to the stochastic note generation is attempted here, by increasing or decreasing the loudness and brightness with one-third probability each every *N* seconds.

The final (*rhythm*) improvement to the stochastic music generation is found by looking at the *rhythmogram* in fig 3. Three things of importance to rhythm are observed in fig 3, and in additional observations of the rhythmograms of the 50 songs of [10]. First, the tempo may or may not drift up to perhaps 10%, secondly, there seems to be short passages of perhaps 10 seconds in which the rhythm is lost, i.e. there is no clear repetition rate in the instruments of the music. Finally, there is often inserted another rhythm of another rate.

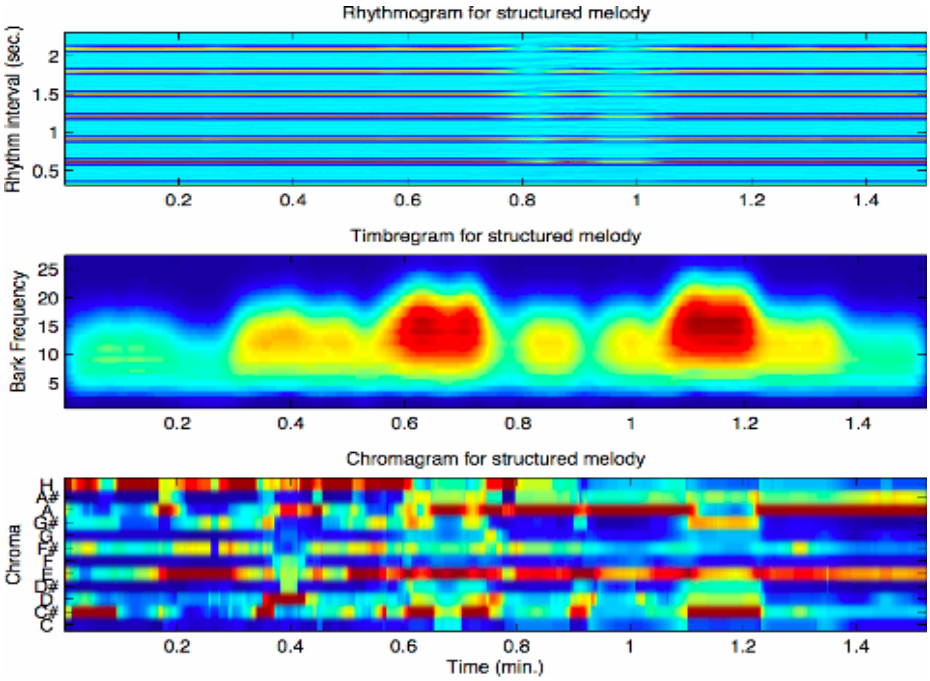

Fig. 7. Rhythmogram, timbregram and chromagram for a test signal with structural changes

The tempo drift is regenerated by inserting a pause with varying length after each note. The length of the pause is governed by a Brownian noise (integrated white noise) whose rate of change decides the tempo drift. As it does not govern the structure of the music, the tempo drift is not included here.

The short passages of arhythmic behavior can be recreated through changing the length of the notes in short passages. This is done by adding a pause of random length for short passages of approximately 10 seconds. Finally, the subtle change of rhythm is not attempted to be modeled here, as it necessitates a rhythm model. Such a rhythm model, while important, is not part of the present work.

As an example of a note-based music with the proposed structural improvement, with regards to rhythm, timbre and chroma, a short melody with these improvements

has been created. The *rhythmogram*, *timbregram*, and *chromagram* of this sound-structure are shown in fig 7.

In an informal listening experiment performed using the authors mainly, the structural change on the chroma is the least perceptible. This is probably related to the stochastic nature of the note generation. The variation of loudness and brightness that induces structure in the timbre, seemingly renders interest and pleasure to the listening experience. The inclusion of arhythmicity that creates structure in the rhythm here dramatically breaks the listening continuity. Overall, by the simple melody and lack of rhythm by the recurrent notes of equal length, it is of course far from complex music, but we deem the structural inclusions promising.

5 Conclusion

In this paper we have compared a human listening strategy with a computational strategy. We have argued that the human tendency for organizing event structures in coherent sections or "super-chunks", with a uniform internal structure and with contrasting features between sections, can be explored in computer based feature extraction. The *rhythmogram*, *timbregram* and *chromagram*, presented here, yield results that can be compared to the 'actual' analysis of the two pieces used.

The two approaches seem to be reasonably compatible. It would therefore be interesting to see if we can implement some of the methods used in analyzing music into ways of improving computerbased music generation. A possible, simple, model for changing the music in order to obtain structural changes is presented here. Based on a stochastic note generator, changes to the possible notes in the current alphabet, the loudness and brightness of the notes and the interval between notes recreate music with structural elements similar to the music examples.

When comparing the two approaches we are in a way comparing human and computerbased cognition. Without entering into a philosophical discussion of the similarities and differences between human beings and computers, we would like to point out that the current study takes advantage of one obvious distinction between the two. A computational approach is basically a bottom-up approach, built on discrete events (in this case notes), while the human approach combines a bottom-up approach with a top-down approach, as seen in the organization of perceptual information is chunks and super-chunks. Further studies will not only lead to a 'naturalization' of computer generated music, but could also enhance our understanding of human cognition.

References

1. Trevarthen, C.: Musicality and the Intrinsic Motor Pulse. Musicae Scientiae, 155–211 (special issue, 2000)
2. Snyder, B.: Music and Memory. An Introduction. The MIT Press, Cambridge (2000)
3. Godøy, R.I.: Gestural-Sonorous Objects: embodied extensions of Schaeffer's conceptual apparatus. Organised Sound 11(2), 149–157 (2006)
4. Kühl, O.: Musical Semantics. Peter Lang, Bern (2007)
5. Walters, J.: Jamie Walters, Atlantic (1994)

6. Barenboim, D.: Mozart: Complete Piano Sonatas and Variations. EMI Classics (1991)
7. Gordon, J.W.: The perceptual attack time of musical tones. J. Acoust. Soc. Am. 82(2) (July 1987)
8. Jensen, K.: A Causal Rhythm Grouping. In: Wiil, U.K. (ed.) CMMR 2004. LNCS, vol. 3310, pp. 83–95. Springer, Heidelberg (2005)
9. Hermansky, H.: Perceptual linear predictive (plp) analysis of speech. J. Acoust. Soc. Am. 87(4), 131–134 (1990)
10. Jensen, K.: Multiple scale music segmentation using rhythm, timbre and harmony. EURASIP Journal on Applied Signal Processing, Special issue on Music Information Retrieval Based on Signal Processing (2006)
11. Sekey, A., Hanson, B.A.: Improved 1-bark bandwidth auditory filter. J. Acoust. Soc. Am. 75(6) (1984)
12. Eckmann, J.P., Kamphorst, S.O., Ruelle, D.: Recurrence plots of dynamical systems. Europhys. Lett. 4, 973–977 (1987)
13. Foote, J.: Automatic Audio Segmentation using a Measure of Audio Novelty. In: Foote, J. (ed.) Proceedings of IEEE International Conference on Multimedia and Expo., vol. 1, pp. 452–455 (2000)
14. Jensen, K., Xu, J., Zachariasen, M.: Rhythm-based segmentation of Popular Chinese Music. In: Proceedings of the ISMIR, London, UK, pp. 374–380 (2005)
15. Gray, R.M., Davidson, L.D.: An introduction to statistical signal processing. Cambridge University Press, Cambridge (2004)