MULTI-FEATURE MODELING OF PULSE CLARITY: DESIGN, VALIDATION AND OPTIMIZATION

Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, Jose Fornari
Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä
<first.last>@campus.jyu.fi

ABSTRACT

Pulse clarity is considered as a high-level musical dimension that conveys how easily in a given musical piece, or a particular moment during that piece, listeners can perceive the underlying rhythmic or metrical pulsation. The objective of this study is to establish a composite model explaining pulse clarity judgments from the analysis of audio recordings. A dozen of descriptors have been designed, some of them dedicated to low-level characterizations of the onset detection curve, whereas the major part concentrates on descriptions of the periodicities developed throughout the temporal evolution of music. A high number of variants have been derived from the systematic exploration of alternative methods proposed in the literature on onset detection curve estimation. To evaluate the pulse clarity model and select the best predictors, 25 participants have rated the pulse clarity of one hundred excerpts from movie soundtracks. The mapping between the model predictions and the ratings was carried out via regressions. Nearly a half of listeners' rating variance can be explained via a combination of periodicity-based factors.

1 INTRODUCTION

This study is focused on one particular high-level dimension that may contribute to the subjective appreciation of music: namely *pulse clarity*, which conveys how easily listeners can perceive the underlying pulsation in music. This characterization of music seems to play an important role in musical genre recognition in particular, allowing a finer discrimination between genres that present similar average tempo, but that differ in the degree of emergence of the main pulsation over the rhythmic texture.

The notion of pulse clarity is considered in this study as a subjective measure that listeners were asked to rate whilst listening to a given set of musical excerpts. The aim is to model these behavioural responses using signal processing and statistical methods. An understanding of pulse clarity requires the precise determination of *what* is pulsed, and *how* it is pulsed. First of all, the temporal evolution of the music to be studied is usually described with a curve – denominated throughout the paper *onset detection curve* – where peaks indicate important events (considered as pulses, note onsets, etc.) that will contribute to the evocation of pulsation. In the proposed framework, the estimation of these primary representations is based on a compilation of state-of-the-art research in this area, enumerated in section 2. In a second step, the characterization of the pulse clarity is estimated through a description of the onset detection curve, either focused on local configurations (section 3), or describing the presence of periodicities (section 4). The objective of the experiment, described in section 5, is to select the best combination of predictors articulating primary representations and secondary descriptors, and correlating optimally with listeners' judgements.

The computational model and the statistical mapping have been designed using *MIRtoolbox* [11]. The resulting pulse clarity model, the onset detection estimators, and the statistical routines used for the mapping, have been integrated in the new version of *MIRtoolbox*, as mentioned in section 6.

2 COMPUTING THE ONSET DETECTION FUNCTION

In the analysis presented in this paper, several models for onset or beat detection and/or tempo estimation have been partially integrated into one single framework. Beats are considered as prominent energy-based onset locations, but more subtle onset positions (such as harmonic changes) might contribute to the global rhythmic organisation as well.

A simple strategy consists in computing the root-mean-square (RMS) energy of each successive frame of the signal ("rms" in figure 1). More generally, the estimation of the onset positions is based on a decomposition of the audio waveform along distinct frequency regions.

- This decomposition can be performed using a bank of filters ("filterbank"), featuring between six [14], and more than twenty bands [9]. Filterbanks used in the models are Gammatone ("Gamm." in table 1) and two sets of non-overlapping filters ("Scheirer" [14] and "Klapuri" [9]). The envelope is extracted from each band through signal rectification, low-pass filtering and down-sampling. The low-pass filtering ("LPF") is implemented using either a simple auto-regressive fil-

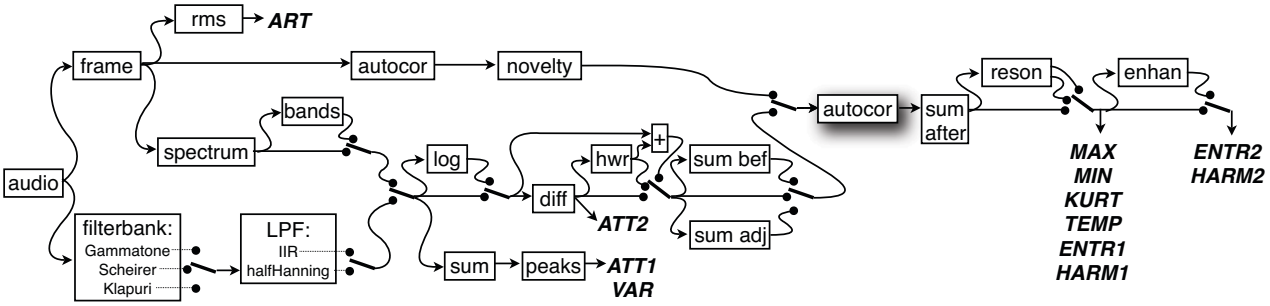

Figure 1. Flowchart of operators of the compound pulse clarity model, where options are indicated by switches.

ter ("IIR") or a convolution with a half-Hanning window ("halfHanning") [14, 9].

- Another method consists in computing a spectrogram ("spectrum") and reassigning the frequency ranges into a limited number of critical bands ("bands") [10]. The frame-by-frame succession of energy along each separate band, usually resampled to a higher rate, yields envelopes.

Important note onsets and rhythmical beats are characterised by significant rises of amplitude in the envelope. In order to emphasize those changes, the envelope is differentiated ("diff"). Differentiation of the logarithm ("log") of the envelope has also been advocated [9, 10]. The differentiated envelope can be subsequently half-wave rectified ("hwr") in order to focus on the increase of energy only. The half-wave rectified differentiated envelope can be summed ("+" in figure 1) with the non-differentiated envelope, using a specific λ weight fixed here to the value .8 proposed in [10] ("λ=.8" in tables 1 and 2).

Onset detection based on spectral flux ("flux" in table 1) [1, 2] – i.e. the estimation of spectral distance between successive frames – corresponds to the same envelope differentiation method ("diff") computed using the spectrogram approach ("spectrum"), but usually without reassignment of the frequency ranges into bands. The distances are hence computed for each frequency bin separately, and followed by a summation along the channels. Focus on increase of energy, where only the positive spectral differences between frames are summed, corresponds to the use of half-wave rectification. The computation can be performed in the complex domain in order to include phase information [1] [2].

Another method consists in computing distances not only between strictly successive frames, but also between all frames in a temporal neighbourhood of pre-specified width [3]. Inter-frame distances [2] are stored into a similarity matrix, and a "novelty" curve is computed by means of a convolution along the main diagonal of the similarity matrix with a Gaussian checkerboard kernel [8]. Intuitively, the novelty curve indicates the positions of transitions along the temporal evolution of the spectral distribution. We notice in particular that the use of novelty for multi-pitch extraction [16] leads to particular good results when estimating onsets from violin solos (see Figure 2), where high variability in pitch and energy due to vibrato makes it difficult to detect the note changes using strategies based on envelope extraction or spectral flux only.

[1] This last option, although available in *MIRtoolbox*, has not been integrated into the general pulse clarity framework yet and is therefore not taken into account in the statistical mapping presented in this paper.

[2] In our model, this method is applied to frame-decomposed autocorrelation ("autocor").

3 NON-PERIODIC CHARACTERIZATIONS OF THE ONSET DETECTION CURVE

Some characterizations of the pulse clarity might be estimated from general characteristics of the onset detection curve that do not relate to periodicity.

3.1 Articulation

Articulation, describing musical performances in terms of *staccato* or *legato*, may have an influence in the appreciation of pulse clarity. One candidate description of articulation is based on Average Silence Ratio (ASR), indicating the percentage of frames that have an RMS energy significantly lower than the mean RMS energy of all frames [7]. The ASR is similar to the low-energy rate [6], except the use of a different energy threshold: the ASR is meant to characterize significantly silent frames. This articulation variable has been integrated in our model, corresponding to predictor "*ART*" in Figure 1.

3.2 Attack characterization

Characteristics related to the attack phase of the notes can be obtained from the amplitude envelope of the signal.

- Local maxima of the amplitude envelope can be considered as ending positions of the related attack phases. A complete determination of each attack phase requires therefore an estimation of the starting position,

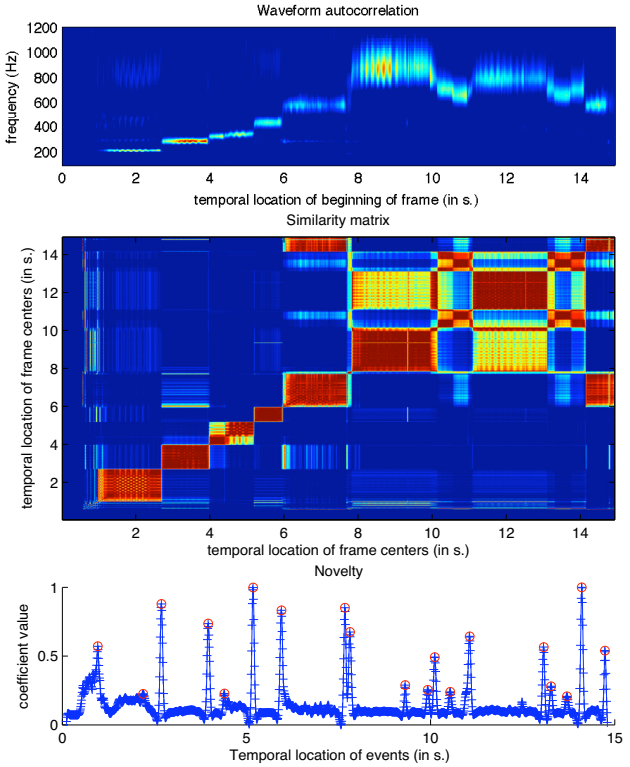

Figure 2. Analysis of a violin solo (without accompaniment). From top to bottom: 1. Frame-decomposed generalized and enhanced autocorrelation function [16] computed from the audio waveform; 2. Similarity matrix measured between the frames of the previous representation; 3. Novelty curve [8] estimated along the diagonal of the similarity matrix with onset detection (circles).

through an extraction of the preceding local minima using an appropriate smoothed version of the energy curve. The main slope of the attack phases [13] is considered as one possible factor (called "*ATT1*") for the prediction of pulse clarity.

- Alternatively, attack sharpness can be directly collected from the local maxima of the temporal derivative of the amplitude envelope ("*ATT2*") [10].

Finally, a variability factor "*VAR*" sums the amplitude difference between successive local extrema of the onset detection curve.

4 PERIODIC CHARACTERIZATION OF PULSE CLARITY

Besides local characterizations of onset detection curves, pulse clarity seems to relate more specifically to the degree of periodicity exhibited in these temporal representations.

4.1 Pulsation estimation

The periodicity of the onset curve can be assessed via autocorrelation ("autocor") [5]. If the onset curve is decomposed into several channels, as is generally the case for amplitude envelopes, the autocorrelation can be computed either in each channel separately, and summed afterwards ("sum after"), or it can be computed from the summation of the onset curves ("sum bef."). A more refined method consists in summing adjacent channels into a lower number of wider band ("sum adj."), on each of which is computed the autocorrelation, further summed afterwards ("sum after") [10].

Peaks indicate the most probable periodicities. In order to model the perception of musical pulses, most perceptually salient periodicities are emphasized by multiplying the autocorrelation function with a resonance function ("reson."). Two resonance curve have been considered, one presented in [15] ("reson1" in table 1), and a new curve developed for this study ("reson2"). In order to improve the results, redundant harmonics in the autocorrelation curve can be reduced by using an enhancement method ("enhan.") [16].

4.2 Previous work: Beat strength

One previous study on the dimension of pulse clarity [17] – where it is termed *beat strength* – is based on the computation of the autocorrelation function of the onset detection curve decomposed into frames. The three best periodicities are extracted. These periodicities – or more precisely, their related autocorrelation coefficients – are collected into a histogram. From the histogram, two estimations of beat strength are proposed: the *SUM* measure sums all the bins of the histogram, whereas the *PEAK* measure divides the maximum value to the main amplitude.

This approach is therefore aimed at understanding the global metrical aspect of an extensive musical piece. Our study, on the contrary, is focused on an understanding of the short-term characteristics of rhythmical pulse. Indeed, even musical excerpts as short as five second long can easily convey to the listeners various degrees of rhythmicity. The excerpts used in the experiments presented in next section are too short to be properly analyzed using the beat strength method.

4.3 Statistical description of the autocorrelation curve

Contrary to the beat strength strategy, our proposed approach is focused on the analysis of the autocorrelation function itself and attempts to extract from it any information related to the dominance of the pulsation.

- The most evident descriptor is the amplitude of the main peak ("*MAX*"), i.e., the global maximum of the curve. The maximum at the origin of the autocorrelation curve is used as a reference in order to normal-

ize the autocorrelation function. In this way, the actual values shown in the autocorrelation function correspond uniquely to periodic repetitions, and are not influenced by the global intensity of the total signal. The global maximum is extracted within a frequency range corresponding to perceptible rhythmic periodicities, i.e. for the range of tempi between 40 and 200 BPM.

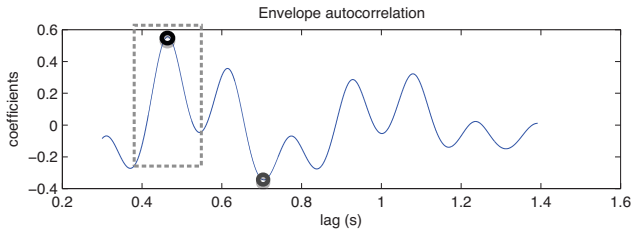

Figure 3. From the autocorrelation curve is extracted, among other features, the global maximum (black circle, *MAX*), the global minimum (grey circle, *MIN*), and the kurtosis of the lobe containing the main peak (dashed frame, *KURT*).

- The global minimum ("*MIN*") gives another aspect of the importance of the main pulsation. The motivation for including this measure lies in the fact that for periodic stimuli with a mean of zero the autocorrelation function shows minima with negative values, whereas for non-periodic stimuli this does not hold true.

- Another way of describing the clarity of a rhythmic pulsation consists in assessing whether the main pulsation is related to a very precise and stable periodicity, or if on the contrary the pulsation slightly oscillates around a range of possible periodicities. We propose to evaluate this characteristic through a direct observation of the autocorrelation function. In the first case, if the periodicity remains clear and stable, the autocorrelation function should display a clear peak at the corresponding periodicity, with significantly sharp slopes. In the second and opposite case, if the periodicity fluctuates, the peak should present far less sharpness and the slopes should be more gradual. This characteristic can be estimated by computing the kurtosis of the lobe of the autocorrelation function containing the major peak. The kurtosis, or more precisely the excess kurtosis of the main peak ("*KURT*"), returns a value close to zero if the peak resembles a Gaussian. Higher values of excess kurtosis correspond to higher sharpness of the peak.

- The entropy of the autocorrelation function ("*ENTR1*" for non-enhanced and "*ENTR2*" for enhanced autocorrelation, as mentioned in section 4.1) characterizes the simplicity of the function and provides in particular a measure of the peakiness of the function. This measure can be used to discriminate periodic and non-periodic signals. In particular, signals exhibiting periodic behaviour tend to have autocorrelation functions with clearer peaks and thus lower entropy than non-periodic ones.

- Another hypothesis is that the faster a tempo ("*TEMP*", located at the global maximum in the autocorrelation function) is, the more clearly it is perceived by the listeners. This conjecture is based on the fact that fast tempi imply a higher density of beats, supporting hence the metrical background.

4.4 Harmonic relations between pulsations

The clarity of a pulse seems to decrease if pulsations with no harmonic relations coexist. We propose to formalize this idea as follows. First a certain number N of peaks [3] are selected from the autocorrelation curve. Let the list of peak lags be $P = \{l_i\}_{i\in[0,N]}$, and let the first peak l_0 be related to the main pulsation. The list of peak amplitudes is $\{r(l_i)\}_{i\in[0,N]}$.

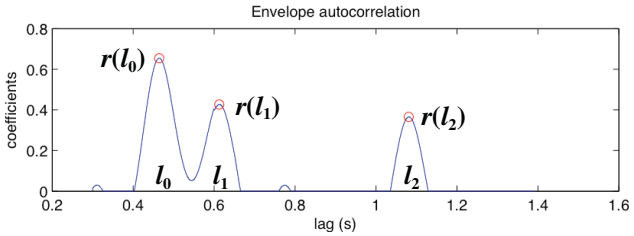

Figure 4. Peaks extracted from the enhanced autocorrelation function, with lags l_i and autocorrelation coefficient $r(l_i)$.

A peak will be inharmonic if the remainder of the euclidian division of its lag l_i with the lag of the main peak l_0 (and the inverted division as well) is significantly high. This defines the set of inharmonic peaks $\overline{H}$:

$$\overline{H} = \left\{ i \in [0,N] \;\middle|\; \begin{array}{ll} l_i \in [\alpha l_0, (1-\alpha) l_0] & (\mathrm{mod}\ l_0) \\ l_0 \in [\alpha l_i, (1-\alpha) l_i] & (\mathrm{mod}\ l_i) \end{array} \right\}$$

where α is a constant tuned to 0.15 in our implementation.

The degree of harmonicity is thus decreased by the cumulation of the autocorrelation coefficients related to the inharmonic peaks:

$$HARM = \exp\left(-\frac{1}{\beta} \frac{\sum_{i\in\overline{H}} r(l_i)}{r(l_0)} \right)$$

where β is another constant, initially tuned [4] to 4.

[3] By default all local maxima showing sufficient contrasts with respect to their adjacent local minima are selected.

[4] As explained in the next section, an automated normalization of the

5 MAPPING MODEL PREDICTIONS TO LISTENERS' RATINGS

The whole set of pulse clarity predictors, as described in the previous sections, has been computed using various methods for estimation of the onset detection curve [5] . In order to assess the validity of the models and select the best predictors, a listening experiment was carried out. From an initial database of 360 short excerpts of movie soundtracks, of 15 to 30 second length each, 100 five-second excerpts were selected, so that the chosen samples qualitatively cover a large range of pulse clarity (and also tonal clarity, another high-level feature studied in our research project). For instance, pulsation might be absent, ambiguous, or on the contrary clear or even excessively steady. The selection has been performed intuitively, by ear, but also with the support of a computational analysis of the database based on a first version of the harmonicity-based pulse clarity model.

25 musically trained participants were asked to rate the clarity of the beat for each of one hundred 5-second excerpts, on a nine-level scale whose extremities were labeled "unclear" and "clear", using a computer interface that randomized the excerpt orders individually [12]. These ratings were considerably homogenous (Cronbach alpha of 0.971) and therefore the mean ratings will be utilized in the following analysis.

Table 1. Best factors correlating with pulse clarity ratings, in decreasing order of correlation r with the ratings. Factor with cross-correlation κ exceeding .6 have been removed.

var	r	κ	parameters
MIN	.59		Klapuri, halfHanning, log, hwr, sum bef., reson1
KURT	.42	.55	Scheirer, IIR, sum aft.
HARM1	.40	.53	Scheirer, IIR, log, hwr, sum aft.
ENTR2	-.4	.54	Klapuri, IIR, log, hwr(λ=.8), sum bef., reson2
MIN	.40	.58	flux, reson1

The best factors correlating with the ratings are indicated in table 1. The best predictor is the global minimum of the autocorrelation function, with a correlation r of 0.59 with the ratings. Hence one simple description of the autocorrelation curve is able to explain already $r^2 = 36$ % of the variance of the listeners' ratings. For the following variables, κ indicates the highest cross-correlation with any factor of better r value. A low κ value would indicate a good independence of the related factor, with respect to the other factors considered as better predictors. Here however, the cross-correlation is quite high, with $\kappa > .5$. However, a stepwise regression between the ratings and the best predictors, as indicated in table 2, shows that a a linear combination of some of the best predictors enables to explain nearly half (47%) of the variability of listeners' ratings. Yet 53% of the variability remains to be explained...

Table 2. Result of stepwise regression between pulse clarity ratings and best predictors, with accumulated adjusted variance r^2 and standardized β coefficients.

step	var	r^2	β	parameters
1	*MIN*	.36	.97	Klapuri, halfHanning, log, hwr, sum bef., reson1
2	*TEMP*	.43	-.5	Gamm., halfHanning, log, hwr, sum aft., reson1
3	*ENTR1*	.47	-.55	Klapuri, IIR, log, hwr(λ=.8), sum bef.

6 *MIRTOOLBOX* 1.2

The whole set of algorithms used in this experiment has been implemented using *MIRtoolbox* [6] [11]: the set of operators available in the version 1.1 of the toolbox have been improved in order to incorporate a part of the onset extraction and tempo estimation approaches presented in this paper. The different paths indicated in the flowchart in figure 1 can be implemented in *MIRtoolbox* in alternative ways:

- The successive operations forming a given process can be called one after the other, and options related to each operator can be specified as arguments. For example,

```
a = miraudio('myfile.wav')
f = mirfilterbank(a,'Scheirer')
e = mirenvelope(f,'HalfHann')
```
etc.

- The whole process can be executed in one single command. For example, the estimation of pulse clarity based on the *MIN* heuristics computed using the implementation in [9] can be called this way:

```
mirpulseclarity('myfile.wav',
                'Min','Klapuri99')
```

distribution of all predictions is carried out before the statistical mapping, rendering the fine tuning of the β constant unnecessary.

[5] Due to the high combinatory of possible configurations, only a part has been computed so far. More complete optimization and validation of the whole framework will be included in the documentation of version 1.2 of *MIRtoolbox*, as explained in the next section.

[6] Available at *http://www.jyu.fi/music/coe/materials/mirtoolbox*

- A linear combination of best predictors, based on the results of the stepwise regression can be used as well. The number of factors to integrate in the model can be specified.

- Multiple paths of the pulse clarity general flowchart can be traversed simultaneously. At the extreme, the complete flowchart, with all the possible alternative switches, can be computed as well. Due to the complexity of such computation [7], optimization mechanisms limit redundant computations.

The routine performing the statistical mapping – between the listeners' ratings and the set of variables computed for the same set of audio recordings – is also available in version 1.2 of *MIRtoolbox*. This routine includes an optimization algorithm that automatically finds optimal Box-Cox transformations [4] of the data, ensuring that their distributions become sufficiently Gaussian, which is a prerequisite for correlation estimation.

7 ACKNOWLEDGEMENTS

This work has been supported by the European Commission (BrainTuning FP6-2004-NEST-PATH-028570), the Academy of Finland (project 119959) and the Center for Advanced Study in the Behavioral Sciences, Stanford University. We are grateful to Tuukka Tervo for running the listening experiment.

8 REFERENCES

[1] Alonso, M., B. David and G. Richard. "Tempo and beat estimation of musical signals", *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[2] Bello, J. P., C. Duxbury, M. Davies and M. Sandler. "On the use of phase and energy for musical onset detection in complex domain", *IEEE Signal Processing. Letters*, 11-6, 553–556, 2004.

[3] Bello, J. P., L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. Sandler. "A tutorial on onset detection in music signals", *Transactions on Speech and Audio Processing.*, 13-5, 1035–1047, 2005.

[4] Box, G. E. P., and D. R. Cox. "An analysis of transformations" *Journal of the Royal Statistical Society. Series B (Methodological)*, 26-2, 211–246, 1964.

[5] Brown, J. C. "Determination of the meter of musical scores by autocorrelation", *Journal of the Acoustical Society of America*, 94-4, 1953–1957, 1993.

[7] In the complete flowchart shown in figure 1, as many as 4383 distinct predictors can be counted.

[6] Burred, J. J., and A. Lerch. "A hierarchical approach to automatic musical genre classification", *Proceedings of the Digital Audio Effects Conference*, London, UK, 2003.

[7] Y. Feng and Y. Zhuang and Y. Pan. "Popular music retrieval by detecting mood", *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, 2003.

[8] Foote, J., and M. Cooper. "Media Segmentation using Self-Similarity Decomposition", *Proceedings of SPIE Conference on Storage and Retrieval for Multimedia Databases*, San Jose, CA, 2003.

[9] Klapuri, A. "Sound onset detection by applying psychoacoustic knowledge", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Phoenix, AZ, 1999.

[10] Klapuri, A., A. Eronen and J. Astola. "Analysis of the meter of acoustic musical signals", *IEEE Transactions on Audio, Speech and Langage Processing*, 14-1, 342–355, 2006.

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

[12] Lartillot, O., T. Eerola, P. Toiviainen and J. Fornari. "Multi-feature modeling of pulse clarity from audio", *Proceedings of the International Conference on Music Perception and Cognition*, Sapporo, Japan, 2008.

[13] Peeters, G. "A large set of audio features for sound description (similarity and classification) in the CUIDADO project (version 1.0)", Report, Ircam, 2004.

[14] Scheirer, E. D. "Tempo and beat analysis of acoustic musical signals", *Journal of the Acoustical Society of America*, 103-1, 588–601, 1998.

[15] Toiviainen, P., and J. S. Snyder. "Tapping to Bach: Resonance-based modeling of pulse", *Music Perception*, 21-1, 43–80, 2003.

[16] Tolonen, T., and M. Karjalainen. "A Computationally Efficient Multipitch Analysis Model", *IEEE Transactions on Speech and Audio Processing*, 8-6, 708–716, 2000.

[17] Tzanetakis, G.,G. Essl and P. Cook. "Human perception and computer extraction of musical beat strength", *Proceedings of the Digital Audio Effects Conference*, Hamburg, Germany, 2002.